Departamento de Matemáticas y Computación

Universidad de La Rioja

Curso básico de análisis de datos con Statistica

Febrero de 2010

VISIT…

Curso básico de análisis de datos con Statistica.

A cargo de: Área de Estadística e Investigación Operativa. Departamento de Matemáticas y Computación. Universidad de La Rioja.

Horario: Días 26 y 28 de enero, 2 y 4 de febrero de 2010, de 16:00 a 19:00 en el Aula Informática nº 4 del CCT.

Profesores: Juan Carlos Fillat Ballesteros
Despacho 226 del edificio Vives. juan-carlos.fillat@unirioja.es
Zenaida Hernández Martín
Despacho 227 del edificio Vives. zenaida.hernandez@unirioja.es
David Ortigosa Martínez
Despacho 206 del edificio Vives. david.ortigosa@unirioja.es
Montserrat San Martín Pérez
Despacho 227 del edificio Vives. montse.sanmartin@unirioja.es

Objetivo: El objetivo de este curso es "Establecer los conceptos básicos para realizar un análisis estadístico metodológicamente correcto, utilizando el programa STATISTICA".

Programa: El curso consta de 4 sesiones con los siguientes contenidos:

1. Estadística Descriptiva 5
   1.1. Organización de datos. Creación de un archivo de datos. 5
   1.2. Análisis estadístico unidimensional. 10
   1.3. Creación de variables y gestión de datos. 20
   1.4. Edición de gráficos. 28
2. Inferencia Paramétrica 33
   2.1. Inferencia Estadística 33
   2.2. Intervalos de Confianza 34
   2.3. Contrastes de Hipótesis paramétricos 38
   2.4. Contrastes paramétricos de dos poblaciones 48
3. Inferencia no paramétrica 59
   3.1. Pruebas de homoscedasticidad 59
   3.2. Pruebas de normalidad 62
   3.3. Contrastes no paramétricos para 1 ó 2 muestras 69
   3.4. Tratamiento de datos cualitativos. Contraste de la χ^2 76
4. Análisis de la Varianza y Regresión Lineal 79
   4.1. Análisis de la Varianza (ANOVA) 79
   4.2. Regresión Lineal Simple 87

Sesión 1. Estadística Descriptiva.

1.1 Organización de datos. Creación de un archivo de datos.

Para poder realizar cualquier análisis estadístico debemos disponer de unos datos. Y estos datos corresponden a los valores obtenidos al estudiar determinadas características en los elementos de un conjunto de entes.

Para fijar el lenguaje que utilizaremos, estableceremos los siguientes términos:

- **Población**: es el conjunto de entes (personas, animales o cosas) sobre los que se va a llevar a cabo la investigación estadística.
- **Elemento**: es cada uno de los componentes de la población (pueden ser simples o compuestos)
- **Caracteres**: son las cualidades o rasgos comunes a toda la población y que vamos a estudiar (pueden ser cuantitativos (**variables**), o cualitativos (**atributos**)).

Aunque existe el análisis estadístico de los caracteres cualitativos (se verá en la tercera sesión), cuando se habla de análisis estadístico, generalmente nos referimos al análisis de las características cuantitativas observadas en los elementos de una población.

Por lo tanto, generalmente trabajaremos con variables estadísticas que, atendiendo a los valores que pueden tomar, pueden ser discretas o continuas; y esta diferencia hace que en muchas ocasiones tengan un tratamiento diferente.

Por otra parte, dentro de los atributos (también llamados variables cualitativas), cabe distinguir dos categorías: los atributos que son simples nombres y/o categorías, y los atributos ordinales que además permiten algún tipo de ordenación.

Otra cuestión muy importante, que se debe tener en cuenta antes de realizar un análisis estadístico es qué es lo que queremos o podemos hacer, en función del tamaño de la población objeto de estudio.

- Si la población es pequeña y podemos obtener datos de todos los elementos de la misma, lo que haremos será un análisis descriptivo (Estadística Descriptiva).
- Pero, si la población es muy grande (infinita o tan grande que no podemos abordarla en su totalidad), no nos queda más remedio que tomar una "muestra representativa", analizar dicha muestra y luego estudiar bajo qué condiciones podemos extender los resultados obtenidos con la muestra a toda la población o si podemos inferir algún resultado para la población. En esto consiste la Inferencia Estadística.

Una vez que tenemos claros estos conceptos, vamos a comenzar un análisis estadístico.

Paso 1: Establecemos la población que queremos estudiar.

Paso 2: Determinamos las características que nos interesa analizar de dicha población.

Paso 3: Recogemos los datos.

Paso 4: Comenzamos el análisis de datos.

Para analizar nuestra "masa de datos", vamos a utilizar el programa **Statistica**. 3

El software STATISTICA almacena los datos en unos ficheros que son como hojas de cálculo y que nombra con la extensión "**.sta**".
¡Ojo!, los ficheros *.sta no se usan exclusivamente para almacenar datos susceptibles de un análisis estadístico, sino que en ellos se puede almacenar también otro tipo de información, por lo que debemos tener muy claro cual es el contenido de nuestro fichero.
Supongamos que disponemos de la siguiente información que queremos analizar:

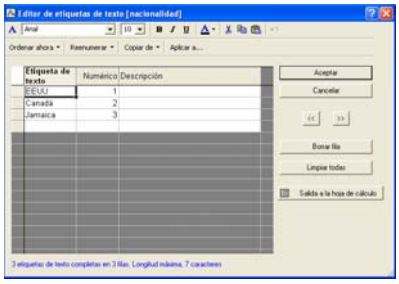
Progresión de los 100 metros lisos masculinos (cronometrados electrónicamente):

Marca	Atleta	Fecha de la marca	Nacionalidad
9.95	Jim Hines	14/10/1968	EEUU
9.93	Calvin Smith	3/07/1983	EEUU
9.92	Carl Lewis	30/08/87	EEUU
9.90	Leroy Burrell	14-06-91	EEUU
9.86	Carl Lewis	25-08-91	EEUU
9.85	Leroy Burrell	06-07-94	EEUU
9.84	Donovan Bailey	29-07-1996	Canadá
9.79	Maurice Greene	16-06-99	EEUU
9.77	Asafa Powell	14-06-2005	Jamaica
9.74	Asafa Powell	9-09-2007	Jamaica
9.72	Usain Bolt	31-05-2008	Jamaica
9.69	Usain Bolt	16-08-2008	Jamaica

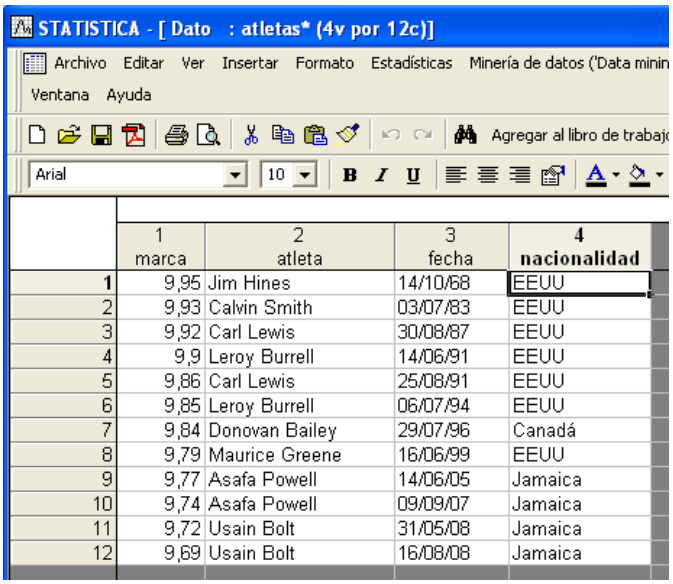
(datos sin actualizar en el 2009)

Para poder hacer cualquier análisis con estos datos debemos crear un fichero de datos adecuado. Tenemos distintas opciones:

a) Crear directamente el fichero con STATISTICA:
   a. Abrimos la aplicación
   b. Archivo/Nuevo/Hoja de cálculo
      i. Especificamos 4 variables y 12 casos (como todas las variables son diferentes, el resto de las características las definiremos más adelante).
      ii. Colocar, como una ventana independiente
   c. Doble clic en el nombre de la primera variable, (vamos a ir definiéndolas todas):
      i. Var 1 --> **Marca**; Número con 2 decimales
      ii. Cambiamos de variable pulsando en ">>"
      iii. Var 2--> **Atleta**; tipo Texto; longitud 20; >>
      iv. Var 3 --> **Fecha marca**; formato Fecha (elige el 6º formato — todas las fechas deben estar en el mismo formato); >>
      v. Var 4 --> **Nacionalidad**; *Etiquetas de texto* (codificamos).

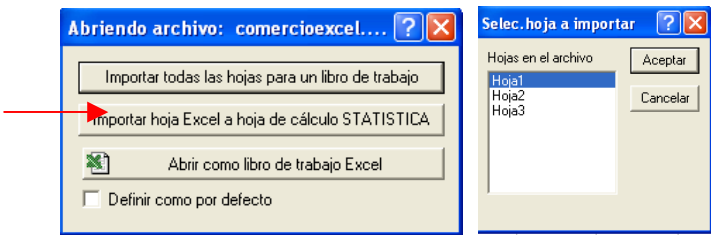


d. Ahora introducimos todos los datos (o los copiamos y pegamos).
e. Guardamos nuestros datos con el nombre **atletas.sta**

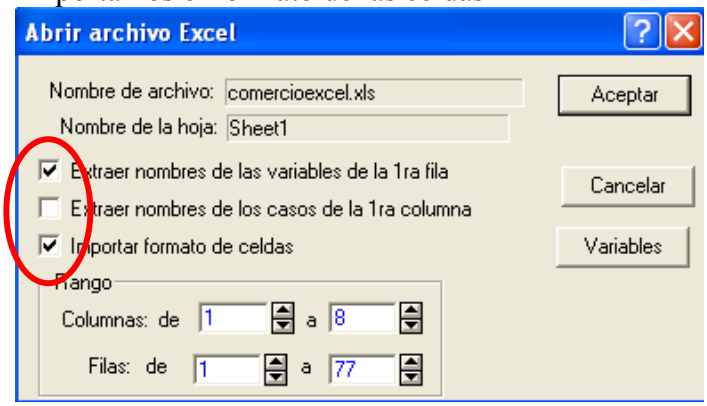

STATISTICA - [Dato : atletas* (4v por 12c)]
Archivo Editar Ver Insertar Formato Estadísticas Minería de datos ('Data minin
Ventana Ayuda
Agregar al libro de trabajo
Arial
10

	1 marca	2 atleta	3 fecha	4 nacionalidad
1	9,95	Jim Hines	14/10/68	EEUU
2	9,93	Calvin Smith	03/07/83	EEUU
3	9,92	Carl Lewis	30/08/87	EEUU
4	9,9	Leroy Burrell	14/06/91	EEUU
5	9,86	Carl Lewis	25/08/91	EEUU
6	9,85	Leroy Burrell	06/07/94	EEUU
7	9,84	Donovan Bailey	29/07/96	Canadá
8	9,79	Maurice Greene	16/06/99	EEUU
9	9,77	Asafa Powell	14/06/05	Jamaica
10	9,74	Asafa Powell	09/09/07	Jamaica
11	9,72	Usain Bolt	31/05/08	Jamaica
12	9,69	Usain Bolt	16/08/08	Jamaica

b) Otra opción es importar el fichero desde otra aplicación (por ejemplo, si tenemos los datos en una hoja de Excel: comercioexcel.xls (***)):
   a. Abrimos la aplicación
   b. Archivo/Abrir.../buscamos el fichero que nos interesa
      i. Si hay más de una hoja del libro con posibles datos a importar, habrá que especificar cuál queremos usar:

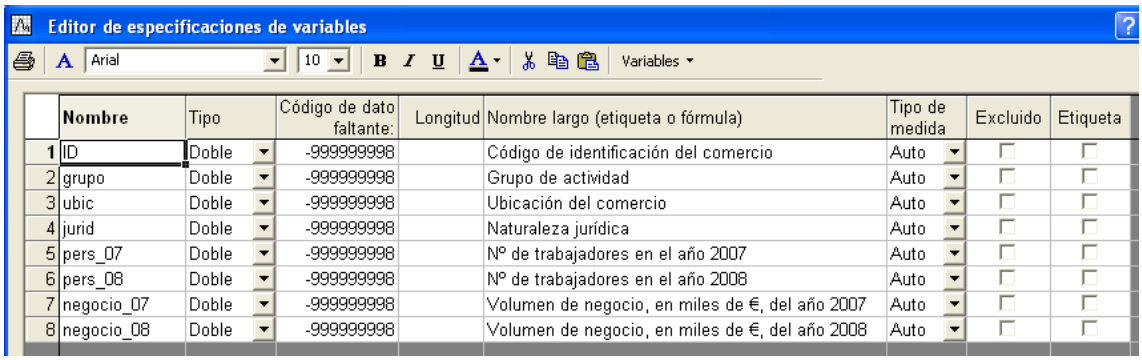


      ii. Especificamos el rango: (por defecto selecciona todas las filas y columnas escritas en la hoja de Excel).
      iii. Importante: señala "Extraer los nombres de la primera fila",
      iv. Importamos el formato de las celdas


Abrir archivo Excel
Nombre de archivo: comercioexcel.xls
Nombre de la hoja: Sheet1
Extraer nombres de las variables de la 1ra fila
Extraer nombres de los casos de la 1ra columna
Importar formato de celdas
Rango
Columnas: de 1 a 8
Filas: de 1 a 77
Aceptar
Cancelar
Variables

   c. Si alguna variable no tiene el formato adecuado lo corregimos, pero ya tenemos nuestro fichero de datos.

Ponemos etiquetas a las variables para indicar cuál es el contenido de las mismas:


Editor de especificaciones de variables

Arial 10 Variables

	Nombre	Tipo	Código de dato faltante:	Longitud	Nombre largo (etiqueta o fórmula)	Tipo de medida	Excluido	Etiqueta
1	ID	Doble	-999999998		Código de identificación del comercio	Auto		
2	grupo	Doble	-999999998		Grupo de actividad	Auto		
3	ubic	Doble	-999999998		Ubicación del comercio	Auto		
4	jurid	Doble	-999999998		Naturaleza jurídica	Auto		
5	pers_07	Doble	-999999998		Nº de trabajadores en el año 2007	Auto		
6	pers_08	Doble	-999999998		Nº de trabajadores en el año 2008	Auto		
7	negocio_07	Doble	-999999998		Volumen de negocio, en miles de €, del año 2007	Auto		
8	negocio_08	Doble	-999999998		Volumen de negocio, en miles de €, del año 2008	Auto		

En un fichero de estas características es conveniente codificar los valores de las variables: grupo, ubicación y naturaleza jurídica. Como puedes comprobar, el sistema lo ha hecho automáticamente al importar los datos (en general lo hace con todas las variables que contienen texto).

NOTA: en Herramientas / Opciones / Hojas de cálculo, se puede modificar número de inicio de los valores que se asignan a las etiquetas de texto.

d. Guardamos nuestros datos con el nombre **comercio.sta** .

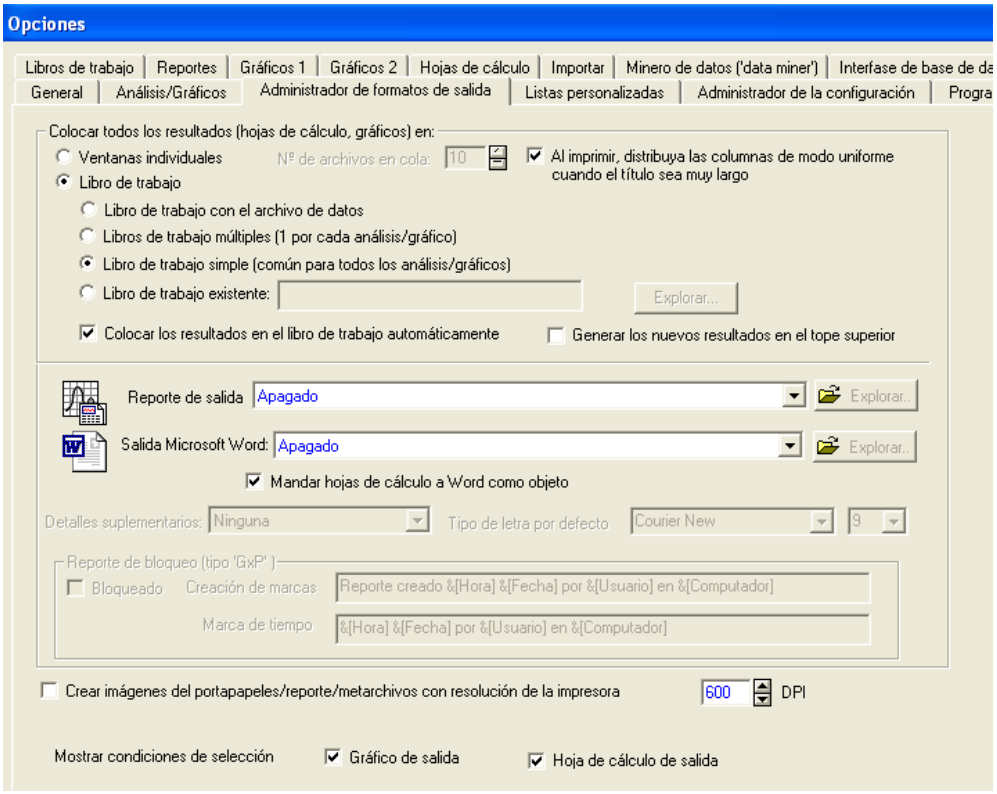
(***) De entre los comercios al por menor de una comunidad autónoma, se toma una muestra.
El archivo **comercioexcel.xls** contiene, para los 76 comercios de la muestra, los datos de las siguientes variables:

ID: Código de identificación del comercio
grupo: Grupo de actividad al que pertenece el comercio (1=”Alimentación”, 2=”Equipamiento personal”, 3=”Equipamiento del hogar”, 4=”Otros”)
ubic: Ubicación del comercio (0=”Dentro de un centro comercial”, 1=”Fuera de cualquier centro comercial”)
jurid: Naturaleza jurídica del comercio (1=”Persona jurídica”, 2=”Sociedad Limitada”, 3=”Sociedad Anónima”, 4=”Otros”).
pers_07: Número de trabajadores en la empresa durante 2007
pers_08: Número de trabajadores en la empresa durante 2008
negocio_07: Volumen de negocio de la empresa en 2007, en miles de €
negocio_08: Volumen de negocio de la empresa en 2008, en miles de €

IMPORTANTE: este fichero se utilizará en las siguientes prácticas.

Configuración de la salida de Statistica.

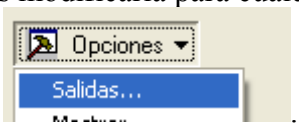
Antes de comenzar con los análisis estadísticos conviene comentar que el programa Statistica nos permite enviar los resultados de cualquier análisis a diferentes documentos de distintos formatos, y es conveniente, antes de empezar, determinar cuál es la salida que nos interesa.
Archivo / Administrador de formatos de salidas...



La opción más habitual, es enviar todos los resultados a un mismo libro de trabajo.

Si estamos realizando un trabajo o queremos elaborar un informe, además del libro de trabajo, tenemos la opción de enviar los resultados a un *Reporte* (documento texto de Statistica) o a un documento de texto de Microsoft *Word*.

La configuración de salida que elijamos permanecerá como opción por defecto hasta que la cambiemos.

También podemos modificarla para cualquier análisis individual en el cuadro de diálogo de dicho análisis: 

.

OBSERVACIÓN: Sea cual sea el análisis o gráfico que estemos realizando, el cuadro de diálogo suele contener opciones de los análisis estadísticos más relacionados, de modo que desde un mismo cuadro de diálogo, podemos realizar un análisis completo, sin tener que estar cambiando de menú.

1.2 Análisis estadístico unidimensional.

Una vez que tenemos nuestro fichero de datos, y suponiendo que tenemos alguna idea del comportamiento de las variables del mismo, una primera aproximación a nuestros datos nos la puede dar un Análisis Estadístico Descriptivo.

Dependiendo de las características de la variable que queramos analizar, podemos utilizar distintos procedimientos.

Si la variable es discreta, y sabemos que en las observaciones aparecen valores "repetidos", tiene sentido realizar una tabla de frecuencias.

Notación: si tenemos una variable estadística X observada sobre una población de tamaño N, dicha variable puede tomar los valores distintos x_1, x_2, ..., x_k, una o más veces. Llamaremos:

Frecuencia absoluta de un determinado valor de la variable x_i (y lo representaremos por $\mathbf{n_i}$) al número de veces que se presenta ese determinado valor x_i.

Nota: $\sum_{i=1}^{k} n_i = N$

Frecuencia relativa de un determinado valor de la variable x_i (y lo representaremos por $\mathbf{f_i}$) a la proporción de veces que se presenta dicho valor, y se calcula como el cociente de su frecuencia absoluta (n_i) y el número total de datos (N).

Frecuencia absoluta acumulada de un determinado valor de la variable x_i (y lo representaremos por $\mathbf{N_i}$) al número de veces que se presenta un valor menor o igual que él y se calcula como la suma de las frecuencias absolutas de todos los valores de la variable menores o iguales que dicho valor x_i.

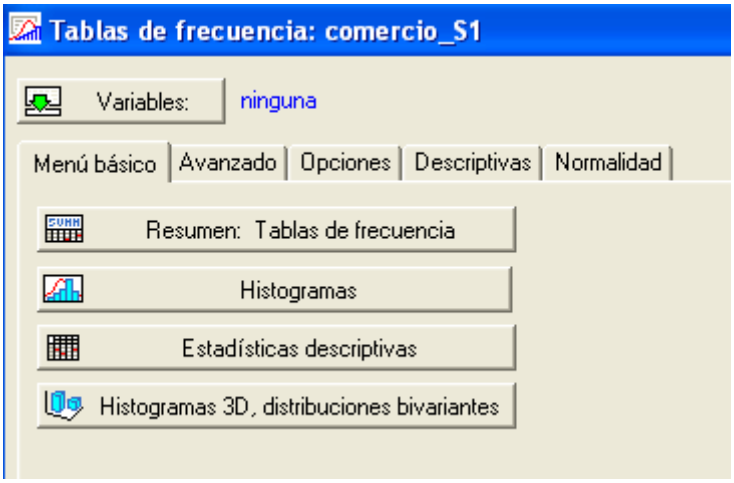
Frecuencia relativa acumulada de un determinado valor de la variable x_i (y lo representaremos por $\mathbf{F_i}$) a la proporción de veces que se presenta un valor menor o igual que él y se calcula como la suma de las frecuencias relativas de todos los valores de la variable menores o iguales que dicho valor x_i.

Nota: En algunas ocasiones los valores se presentan en intervalos (por ejemplo rangos de edad), y en estos casos las frecuencias se refieren, no a los valores concretos de la variable sino, al número de observaciones dentro de cada intervalo. Para hacer los cálculos se utiliza un representante de cada intervalo, que se denomina marca de clase, y suele tomarse el punto medio del mismo.

Cuando en un conjunto de valores observados de una variable, se realizan las operaciones de: Ordenación y Agrupación de los valores que se repiten, (determinación de la frecuencia de cada valor), se obtiene una **tabla estadística de distribución de frecuencias**. A dicho conjunto de operaciones se le denomina Tabulación.

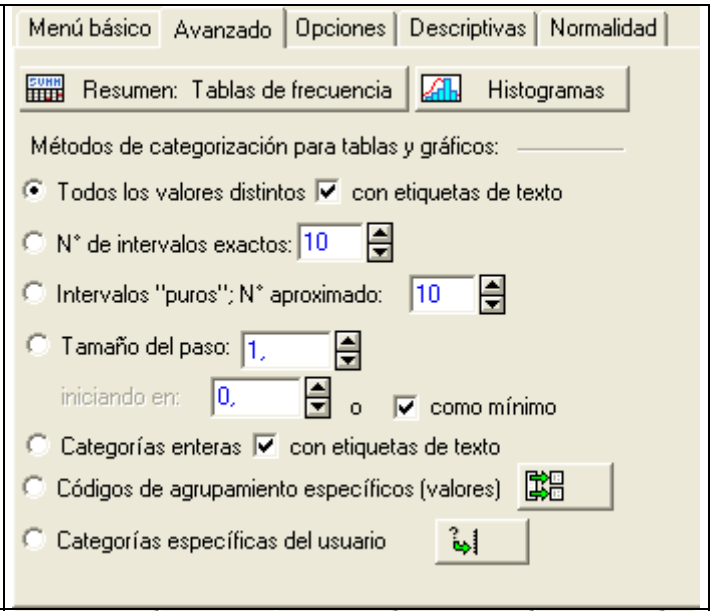
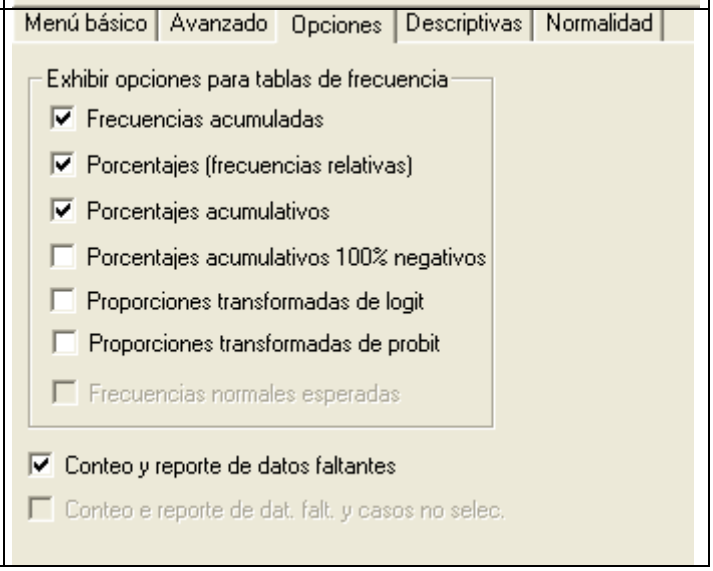
Con Statistica, las tablas de frecuencias se construyen desde el menú:

Estadísticas / Estadísticas/Tablas básicas / Tablas de frecuencia



A continuación se selecciona la variable o variables que queremos analizar.

Además del menú básico en el que se recogen las opciones por defecto, las pestañas que utilizaremos más frecuentemente son:

Avanzado: Para elegir el método de categorización de las variables.	
Opciones: Si queremos modificar las especificaciones de las frecuencias que queremos que aparezcan en las tablas.	

EJERCICIO 1.1.- Construye la tabla de frecuencias de la variable **pers_07**.

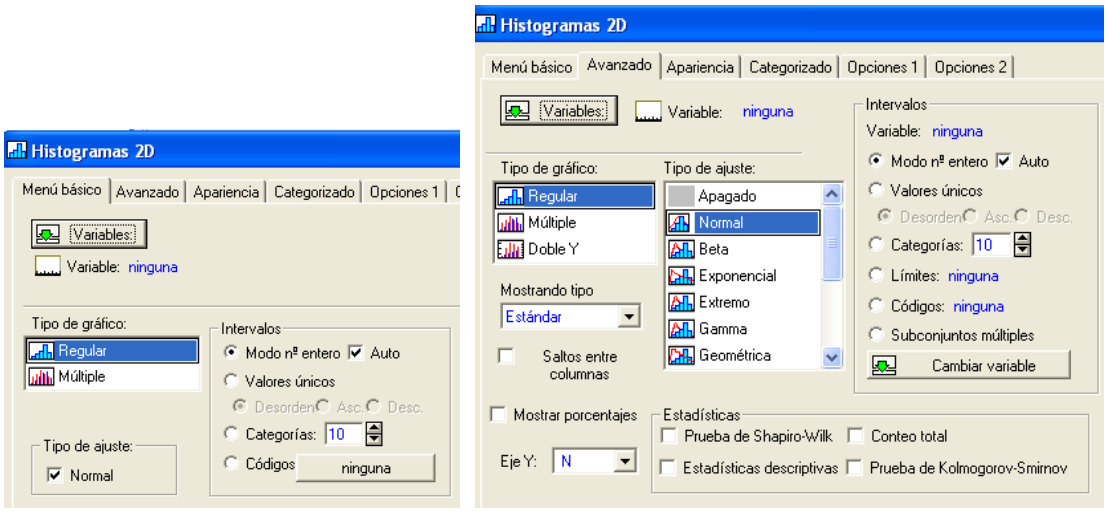
EJERCICIO 1.2.- Construye la tabla de frecuencias de la variable **negocio_08**.
Esta variable es una variable continua, por lo que debemos agrupar en intervalos. Para ello, vamos a la pestaña de Avanzado y seleccionamos, por ejemplo, 6 intervalos.
Observa las diferencias entre Nº de intervalos exactos y aproximados.

Otra forma de presentar la información de una masa de datos es mediante los **gráficos**.

El gráfico de una variable o de una relación entre variables será útil siempre que ponga de manifiesto, de forma sencilla, las características más relevantes de la misma.

En Estadística Descriptiva, los gráficos más sencillos son:

- Los **histogramas**, para representar frecuencias (ya sea de las observaciones que se encuentran dentro de un intervalo como de los valores observados de la variable).



EJERCICIO 1.3.- Construye el histograma del número de empresas según su naturaleza jurídica.

EJERCICIO 1.4.- Construye el histograma del volumen de negocio en el año 2008, (como se puede considerar continua, haz el gráfico con 10 categorías).

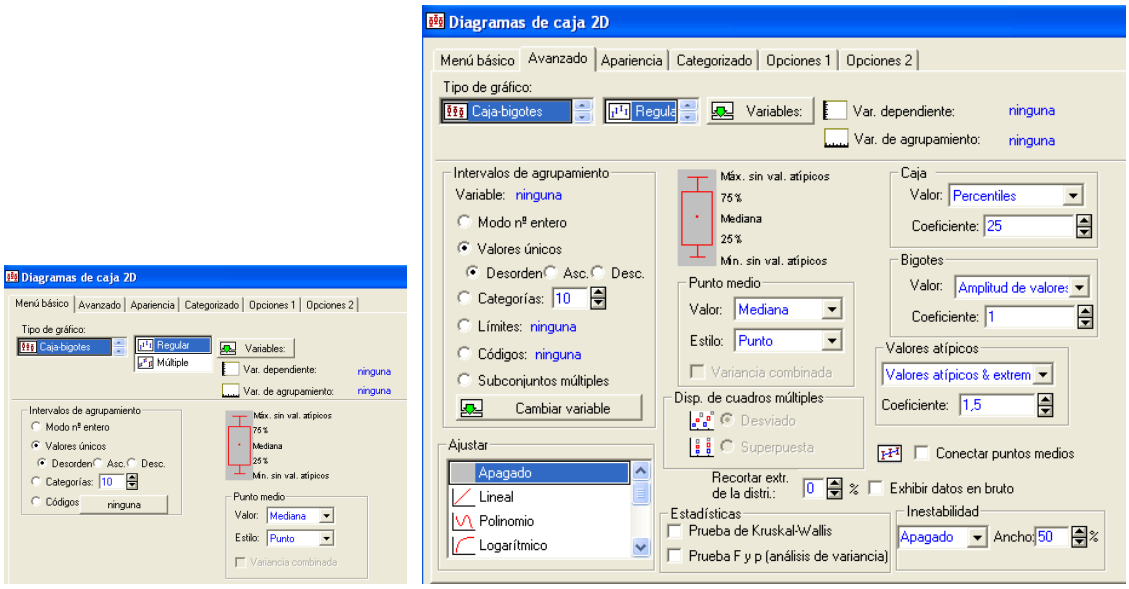
- Los **gráficos de cajas**, para mostrarnos la distribución de los datos observados.

Los gráficos de cajas con bigotes resumen el comportamiento de la variable de la siguiente forma:

1. Una marca central nos indica dónde se encuentra la media.
2. Una caja cuyo tamaño nos indica la media más/menos la cuasidesviación típica.

3. Unos bigotes que nos muestran la media más/menos 1'96* cuasidesviación típica. Esto significa que si nuestros datos siguen una distribución Normal, el 95% de los datos deben estar dentro del rango que abarcan los bigotes.

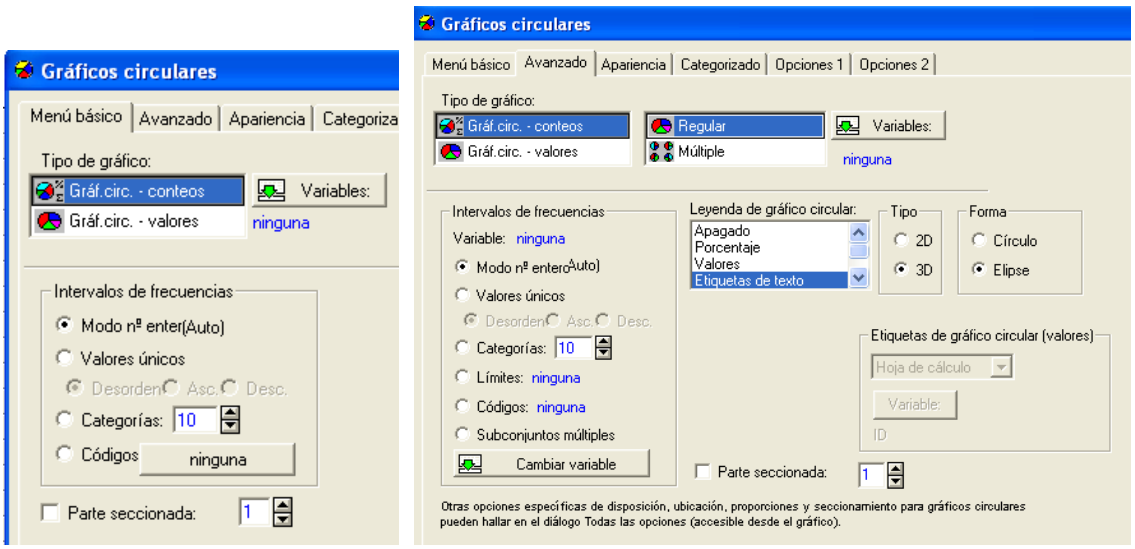
Esta es una de las opciones de los gráficos de cajas, pero también se pueden obtener gráficos de caja con otra información (por ejemplo, basada en los cuartiles).



EJERCICIO 1.5.- Construye el gráfico de caja, basado en los cuartiles, del volumen de negocio en el año 2007.

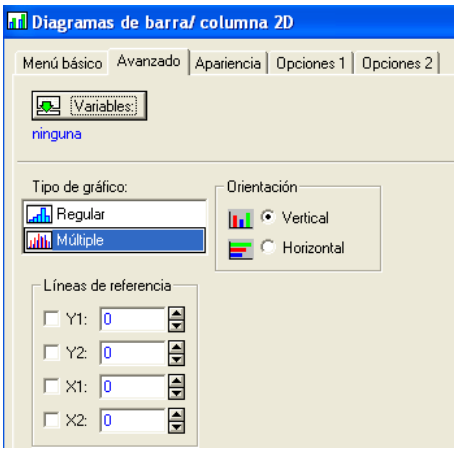
EJERCICIO 1.6.- Realiza un gráfico en el que podamos comparar la distribución, basada en los cuartiles, del volumen de negocio en el año 2008, según la ubicación de las empresas.

- Los **gráficos de sectores**, nos permiten representar frecuencias o valores de la variable, a la vez que podemos visualizar la proporción que representan estas cantidades respecto a la suma total.



EJERCICIO 1.7.- Construye el gráfico de sectores del número de empresas según su naturaleza jurídica.

- Los **gráficos de barras**, nos muestran el valor de la variable, para cada uno de los casos del fichero.



***EJERCICIO 1.8*.-** Realiza el gráfico de barras de la variable **grupo**.

Como se puede observar, este gráfico no nos da un recuento de los distintos valores de la variable (esto lo hacemos con el histograma) sino el valor que toma la variable en cada uno de los casos.

Además de las tablas y los gráficos, para describir el comportamiento de las variables es más preciso calcular una serie de **medidas** (descriptivas) :

a) Medidas de tendencia central y de posición no central.

b) Medidas de dispersión y medidas de forma.

Medidas de tendencia central y de posición no central

Las medidas de posición o promedios, son unos valores alrededor de los cuales se agrupan los valores de la variable, y que nos resumen la posición de la distribución sobre el eje horizontal. Para que un valor pueda ser considerado promedio, se le exige como única condición que esté comprendido entre el mayor y el menor valor de la variable. Existen dos tipos de medidas de posición: las centrales y las no centrales.

De las medidas de posición central, las más utilizadas son: la media aritmética, la mediana y la moda aunque también son muy útiles para calcular determinados promedios la media geométrica y la media armónica.

La media aritmética: se define como la suma de todos los valores de la distribución, dividida por el número total de observaciones:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{k} x_i.n_i = \frac{x_1.n_1 + \cdots + x_k.n_k}{N}$$

Para una distribución discreta no agrupada en intervalos, se define **la mediana**, como el valor de la variable que ocupa el lugar central, supuestos ordenados los valores de menor a mayor.

En una distribución, **la moda** (Mo) se define como "aquel valor de la variable cuya frecuencia no es superada por la frecuencia de ningún otro valor" (puede haber más de una moda).

Para las variables numéricas se pueden calcular estos tres promedios, pero atendiendo al tipo de datos, la moda es el más adecuado para variables nominales o categóricas, la mediana para variables ordinales o categóricas ordinales y la media para las variables estadísticas.
Nota: Si existen valores extremos en la variable, la mediana es mejor medida de la posición central que la media.

Otras medidas, menos habituales, de tendencia central, son:

La **media geométrica**: se define como la raíz N-ésima del producto de los valores de la variable, elevados cada uno de ellos a la potencia indicada por su frecuencia. Es decir:

$$M_G = \sqrt[N]{\prod_{i=1}^{k} x_i^{n_i}} = \sqrt[N]{x_1^{n_1} \ldots x_k^{n_k}}$$

Se utiliza principalmente para promediar porcentajes, tasas, números índices, etc., es decir, en los casos en los que se supone que la variable presenta variaciones acumulativas.

La **media armónica**: Se define como el valor:

$$M_A = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} = \frac{N}{\frac{n_1}{x_1} + \ldots + \frac{n_k}{x_k}}$$

Este promedio se suele utilizar cuando la variable es del tipo: velocidad, rendimiento, etc.

Las medidas de **posición no central o cuantiles**, constituyen una generalización del concepto de la mediana. Así como la mediana divide a la distribución en dos partes, ambas con el mismo número de observaciones, si dividimos la distribución en cuatro partes, cada una de ellas con el mismo número de observaciones, necesitaremos tres valores, que se denominan **cuartiles**.

Análogamente, si dividimos la distribución en diez partes iguales, necesitaremos nueve valores, que se denominan **deciles**. Y si la dividimos en cien partes, los correspondientes noventa y nueve valores, se denominan **percentiles**.

En general, los q-1 valores que dividen a la distribución en q partes iguales (con el mismo número de observaciones) se denominan **cuantiles de orden q**.

Estas medidas se utilizan para estudiar cómo están distribuidos los valores de la variable observada y hacernos una idea de su comportamiento. También nos puede interesar conocer a partir de qué valor de la variable se encuentra un determinado porcentaje de las observaciones mayores o entre qué dos valores se encuentra determinado porcentaje de las observaciones centrales. Para responder a estas preguntas utilizaremos los percentiles.

Cabe recordar que el cálculo de los valores percentiles, aunque se puede realizar con cualquier variable estadística, tiene sentido para variables continuas en las que tenemos muchos valores distintos observados y nos interesa localizar alguna posición dentro de la distribución.

Medidas de dispersión.

Con las medidas de posición o promedios, intentamos sintetizar una tabla de datos. Para evaluar la representatividad de un promedio, necesitamos un indicador que, de alguna forma, nos cuantifique el grado de separación o dispersión de los valores de la variable respecto al promedio en cuestión. Con las medidas de dispersión absolutas, se trata de medir la separación que, por término medio, existe entre los distintos valores de la variable, por lo que serán medidas que vendrán expresadas en la misma clase de unidades que la variable.

Las principales medidas de dispersión absoluta son:

Recorrido (o Rango o Amplitud): se define como la diferencia entre el mayor y el menor valor de la variable. Es decir : $Re = Máx\ x_i - Mín\ x_i = x_k - x_1$

Varianza: la definimos como la media aritmética de los cuadrados de las desviaciones de los valores de la variable respecto a la media aritmética de la distribución. Se representa por S'^2 y es:

$$S'^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N} = \frac{\sum_{i=1}^{k}\left(x_i - \overline{X}\right)^2 .n_i}{N}$$

Desviación típica: o desviación estándar, es igual a la raíz cuadrada de la varianza, con signo positivo. Se representa por S':

$$S' = +\sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N}} = +\sqrt{\frac{\sum_{i=1}^{k}\left(x_i - \overline{X}\right)^2 .n_i}{N}}$$

De todas las medidas de dispersión absoluta, la varianza y su raíz cuadrada, la desviación típica, son las más importantes. Sin embargo, en inferencia son más utilizadas las siguientes:

Cuasivarianza: Es una medida muy similar a la varianza:

$$S^2 = \frac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N-1} = \frac{\sum_{i=1}^{k}\left(x_i - \overline{X}\right)^2 .n_i}{N-1}$$

Cuasidesviación típica: $S = +\sqrt{\dfrac{\sum_{i=1}^{N}\left(x_i - \overline{X}\right)^2}{N-1}} = +\sqrt{\dfrac{\sum_{i=1}^{k}\left(x_i - \overline{X}\right)^2 .n_i}{N-1}}$

Como se puede observar la única diferencia es que las medidas utilizadas en inferencia tienen N-1 en el denominador, en lugar de N.

IMPORTANTE: Cuando en los programas estadísticos se dice varianza y desviación típica, en realidad lo que se calcula es la cuasivarianza y su raíz cuadrada la cuasidesviación típica.

De ahora en adelante las nombraremos indistintamente, pero nos estaremos refiriendo siempre a la cuasivarianza y a la cuasidesviación típica.

Con las medidas de dispersión relativas, se trata de medir la dispersión, con independencia de la clase de unidades en que venga expresada la variable. Estas medidas, permiten comparar la dispersión existente en dos distribuciones, incluso cuando las variables estén expresadas en distinta clase de unidades.

Entre las medidas de dispersión relativa, llamadas también índices de dispersión, tenemos:

Recorrido relativo: Se define como el cociente entre el recorrido de la variable y la media aritmética: $R_r = \dfrac{\mathrm{Re}}{\overline{X}}$

Nos indica el número de veces que el recorrido contiene a la media aritmética.

Coeficiente de variación o índice de dispersión de Pearson: Es el más empleado de los índices de dispersión relativos. Se define como el cociente entre la desviación típica y la media aritmética. $CV = \dfrac{S}{\overline{X}}$

Nos indica el número de veces que la desviación típica contiene a la media aritmética.

Medidas de forma (simetría y curtosis)

Vamos a ver ahora ciertas medidas que nos van a dar una idea de la forma de la distribución, sin necesidad de realizar su representación gráfica.

La distribución NORMAL, cuya representación gráfica es la curva de Gauss, es el modelo de comparación para la simetría y la curtosis de cualquier distribución de frecuencias. Por lo tanto, tiene sentido calcular estas medidas cuando la distribución de la variable objeto de estudio es campaniforme y unimodal.

Las **medidas de simetría** se dirigen a construir un indicador que permita establecer el grado de simetría o asimetría que presenta la distribución, sin necesidad de llevar a cabo su representación gráfica. Diremos que una distribución es simétrica cuando lo es su representación gráfica en coordenadas cartesianas.

Existen distintos coeficientes que nos permiten establecer el grado de asimetría de una variable, y aunque los valores numéricos de los mismos no coincidan, la interpretación sí que es análoga:

- Si el coeficiente de asimetría = **0**, la distribución puede ser simétrica
- Si el coeficiente de asimetría > **0**, la distribución es asimétrica positiva (a la derecha)
- Si el coeficiente de asimetría < **0**, la distribución es asimétrica negativa (a la izquierda)

Observación: toda distribución simétrica tiene nulo el coeficiente de asimetría, pero el recíproco no es cierto, es decir: existen distribuciones asimétricas para las que el índice de asimetría es nulo.

Medidas de apuntamiento o curtosis.

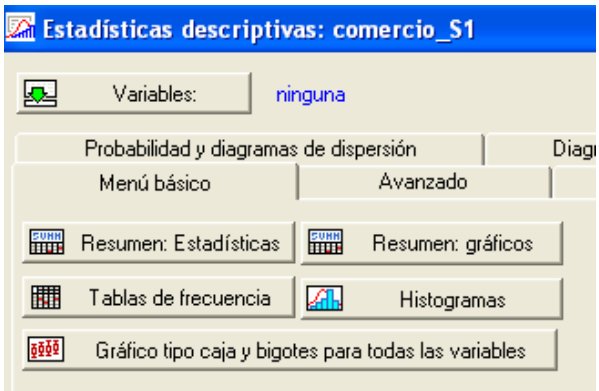
La mayor o menor agrupación de frecuencias alrededor de la media y en la zona central de la distribución, dará lugar a una distribución más o menos apuntada. Existen distintos coeficientes que nos permiten establecer el grado de curtosis de una variable, y aunque los valores numéricos de los mismos no coincidan, la interpretación es análoga:

- Si el coeficiente de curtosis = **0** , la distribución es mesocúrtica (normal).
- Si el coeficiente de curtosis > **0** , la distribución es leptocúrtica (más apuntada que la normal).
- Si el coeficiente de curtosis < **0** , distribución platicúrtica (menos apuntada que la normal).

Cálculos con el Statistica:

Todos los análisis de Estadística Descriptiva (incluyendo las tablas de frecuencias y los gráficos más sencillos), se pueden realizar desde el menú:

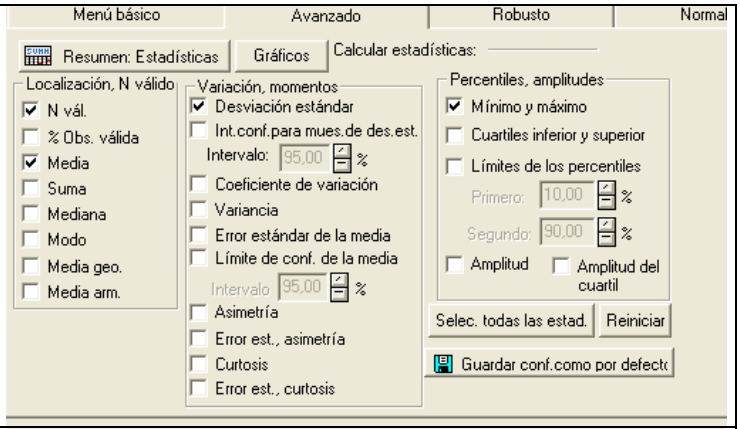
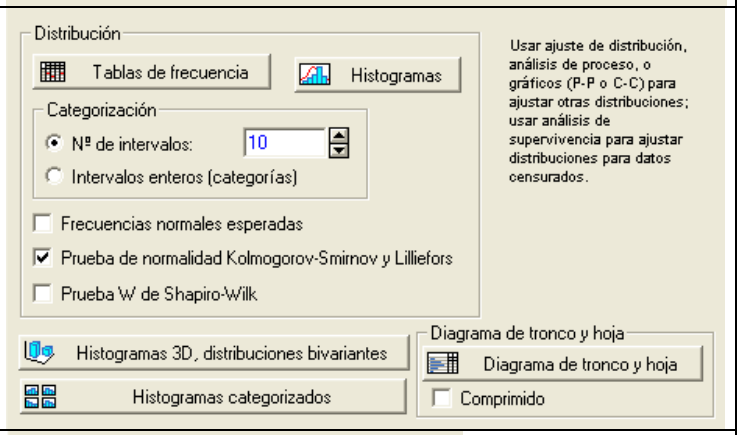
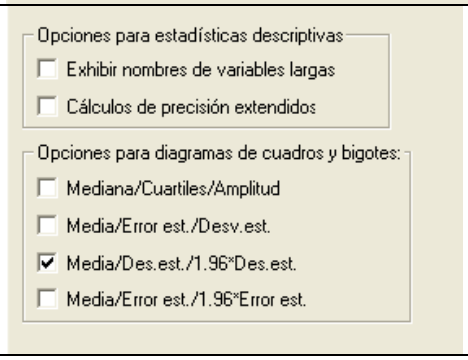
Estadísticas / Estadísticas/Tablas básicas / Estadísticas descriptivas



Como ves en el menú básico se puede acceder no sólo al resumen estadístico sino también a las tablas de frecuencias, los histogramas y los gráficos de caja.

Nota: en este menú, se considera que las variables son continuas, por lo que si la variable es categórica, habrá que especificarlo.

Además del menú básico en el que se recogen las opciones por defecto, las pestañas que utilizaremos más frecuentemente son:

Avanzado: Para elegir los estadísticos descriptivos que nos interesen en nuestro análisis.	
Normalidad: Para especificar si una variable es continua o categórica y para pedir pruebas de normalidad.	
Opciones: Para elegir el tipo de gráfico de caja con bigotes (entre los que tiene el sistema por defecto en este menú)	

EJERCICIO 1.9.- Calcula la media, la mediana y la moda así como los cuartiles y el percentil 82 de las variables **negocio_07** y **negocio_08**.

***EJERCICIO 1.10*.-** Calcula los cuartiles de la variable **negocio_08**.

***EJERCICI 1.11*.-** Calcula la media, la desviación típica y el coeficiente de variación de la variable **negocio_07**.

***EJERCICIO 1.12*.-** Calcula el valor de la variable **negocio_08** que no es superado por el 33% de las observaciones (P_{33}).

***EJERCICIO 1.13*.-** Calcula el valor de la variable **negocio_07** que sólo es superado por el 33% de las observaciones (P_{67}).

***EJERCICIO 1.14*.-** Calcula los valores que encierran el 38% central de las observaciones (P_{31} y P_{69}) para las variables **negocio_07** y **pers_07**.

***EJERCICIO 1.15*.-** Calcula la media aritmética, la media geométrica y la media armónica de las variables **negocio_07 , pers_07 , negocio_08 y pers_08**, y comprueba que se verifica que: $M_A \leq M_G \leq \overline{X}$.

***EJERCICIO 1.16*.-** Estudia la asimetría y la curtosis de las variables **negocio_07** , **pers_07** , **negocio_08** y **pers_08.** Para ello obtén los coeficientes de asimetría y curtosis y los histogramas con curva Normal (comprueba los resultados).

***EJERCICIO 1.17*.-** Obtén las tablas de frecuencias de las variables **grupo** y **pers_08** .

Guarda el libro de trabajo, en el que tienes todos los resultados.

1.3 Creación de variables y gestión de datos

En algunas ocasiones, no sólo tenemos que trabajar con los datos obtenidos en la observación de una población o una muestra, sino que nos interesa calcular nuevas variables como combinación de las anteriores o con valores constantes que nos permitan completar nuestro estudio.

***EJERCICIO 1.18*.-** Supongamos que nos interesa conocer la diferencia entre el volumen de negocio del año 2007 y el año 2008, para cada una de las empresas. Vamos a crear una nueva variable llamada **diferencia**, que contenga estas cantidades:

Añadimos una nueva variable al fichero: Dato /Variables / Agregar... (se puede hacer de muchas formas, pero la más sencilla es hacer doble clic en la zona vacía del fichero de datos).

a. Cuántas variables: 1
b. Después de: negocio_08
c. Nombre: **diferencia**
d. Formato: General
e. Fórmula: =negocio_08-negocio_07
   O bien: =v8-v7

***EJERCICIO 1.19*.-** Supongamos que interesa conocer y analizar el incremento relativo que representa el volumen de negocio en el año 2008 respecto al del 2007 (en tanto por ciento).

Crearemos una variable llamada **variación** que contenga dicho incremento. En este caso, la fórmula a utilizar será:
Fórmula: =100*(negocio_08-negocio_07)/negocio_07
O bien: =100*(v8-v7)/v7
(pon formato: nº con 2 decimales)

NOTA: Si hacemos clic en el botón **Funciones**, veremos la lista completa de las funciones que tiene predefinidas este programa.

Guardamos el fichero de datos (basta con guardar si no queremos cambiarle el nombre).

Recodificar

Recodificar una variable consiste en cambiar sus valores por otros nuevos siguiendo una regla preestablecida (que corresponderá a algún objetivo de nuestro análisis). Por ejemplo, supongamos que en el archivo **comercio** deseamos clasificar las empresas en tres niveles según el volumen de negocio del año 2008, de acuerdo con la siguiente regla:

Si **negocio_08** < P_{25}, entonces la empresa es de nivel **1=”Bajo”**

Si $P_{25} \leq$ **negocio_08** $\leq P_{75}$, entonces la empresa es de nivel **2=”Medio”**

Si P_{75} < **negocio_08**, entonces la empresa es de nivel **3=”Alto”**

Como paso preliminar antes de entrar en la recodificación propiamente dicha es necesario que determinemos los valores de P_{25} (percentil 25 ó primer cuartil) y P_{75} (percentil 75 ó tercer cuartil) para la variable **negocio_08**.

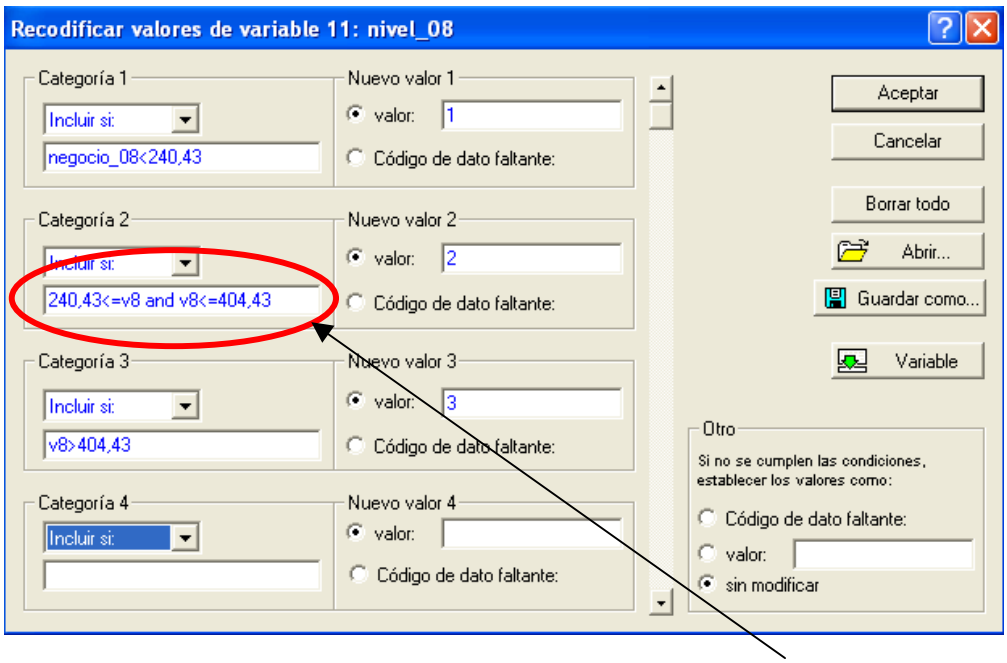
Obtenemos en el libro de trabajo la siguiente información:

	Estadísticas descripti	
Variable	**Inferior Cuartil**	Superior Cuartil
negocio_08	240,4300	404,4300

Por otra parte, y como ya hemos dicho, la recodificación sustituye los valores originales por los valores nuevos, por lo que si no queremos perder la información original, tendremos que crear una nueva variable (puede ser vacía) en la que guardaremos los valores codificados:

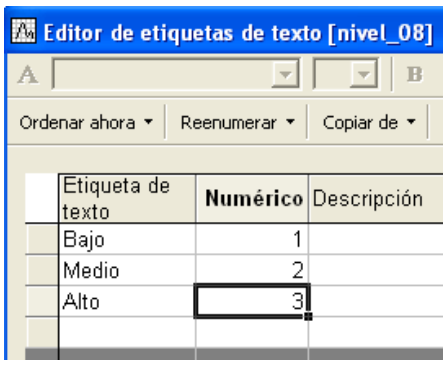
Vars / Agregar... / (**nivel_08**)

Ahora podemos entrar ya en el proceso de recodificación: Dato / Recodificar....



Importante: el sistema no admite expresiones complejas. Hay que descomponerlas en expresiones simples.

Para la nueva variable, podemos crear etiquetas:



Guarda el fichero de datos.

En el archivo **comercio** tenemos ahora la nueva variable **nivel_08** que contiene nuestra recodificación de la variable de partida **negocio_08**. Para esta nueva variable podemos calcular, por ejemplo, su distribución de frecuencias, obteniéndose unos resultados acordes con lo que cabía esperar dado el criterio de recodificación utilizado.

Categoría	Tabla de frecuencia: nivel_08 (comercio_S1)			
	Frecuencia	Acumul. de Frecuencia	%	Acumul. de %
Bajo	18	18	23,68421	23,6842
Medio	38	56	50,00000	73,6842
Alto	18	74	23,68421	97,3684
Faltante	2	76	2,63158	100,0000

Tipificar (Estandarizar)

Sabemos que una variable tipificada es aquella que tiene media igual a cero (variable centrada) y desviación típica igual a 1.

En determinados análisis estadísticos nos interesará que nuestras variables tengan estas características.
Podemos hacerlo "a mano" calculando la media y la desviación típica de la variable correspondiente y luego calculando la variable tipificada, pero no vale la pena el trabajo ya que Statistica nos calcula directamente los valores tipificados de las variables.

Con el menú Dato / Estandarizar , para las variables seleccionadas, el sistema **sustituye** los valores de las variables seleccionadas por sus correspondientes valores tipificados.

Esto está bien si no nos importa perder los valores originales, pero si queremos mantenerlos, previamente habrá que duplicar las variables que queremos tipificar. Por lo tanto el procedimiento es el siguiente:

1) Duplicamos las variables a tipificar creando nuevas variables cuyos valores sean iguales a los de las variables a tipificar.
2) Tipificamos estas variables.

***EJERCICIO 1.20*.-** Obtén los valores tipificados de las variables negocio_07 y negocio_08.

Guarda el fichero de datos.

***EJERCICIO 1.21*.-** Comprueba que efectivamente las nuevas variables tienen media igual a cero y desviación típica igual a 1.

Variable	Estadísticas descriptivas (comercio_S1)				
	N vál.	Media	Mínimo	Máximo	Desv. est.
Z-negocio_07	76	0,000000	-2,11068	2,911664	1,000000
Z-negocio_08	74	0,000000	-2,26360	1,985758	1,000000

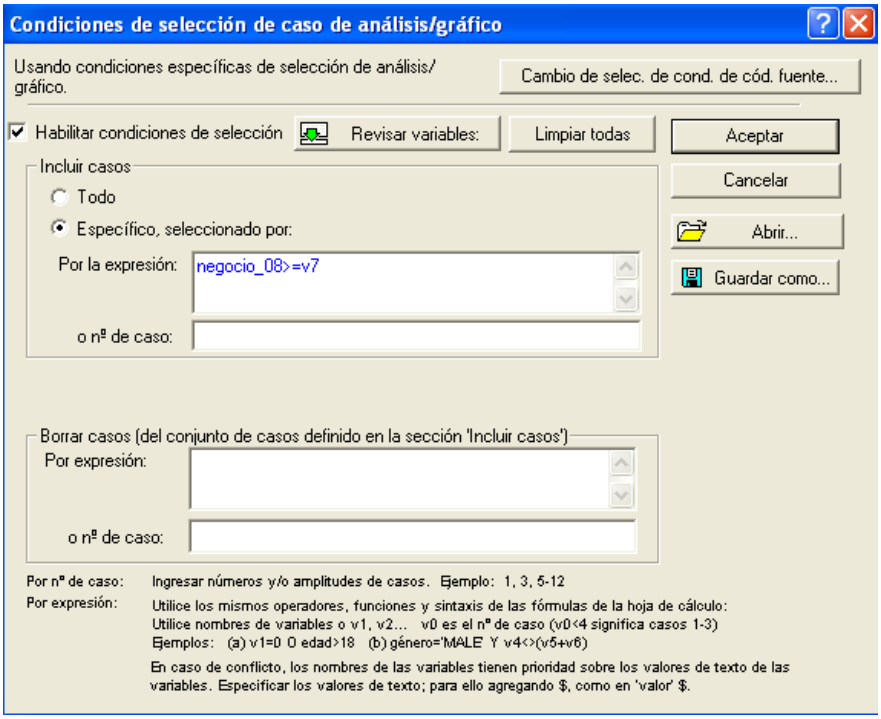
Seleccionar casos

No siempre el análisis estadístico que queremos realizar se referirá al archivo de datos completo. A menudo estaremos interesados en analizar un subconjunto de los casos existentes en el fichero.

Supongamos, por ejemplo, que en el archivo **comercio** deseamos realizar un análisis referido únicamente a las empresas que en el año 2008 han tenido un volumen de negocio mayor o igual que el del año 2007. Debemos "seleccionar" los casos que satisfacen dicha condición, para lo cual procederemos de la siguiente manera:

En cualquiera de los menús, siempre aparece el botón

. Este nos permite, realizar el análisis que queramos, pero sólo para los casos que seleccionemos.
El cuadro de diálogo nos permite seleccionar casos que cumplan una condición, casos sueltos y eliminar algunos de los casos ya seleccionados, con criterios similares:


Condiciones de selección de caso de análisis/gráfico
Usando condiciones específicas de selección de análisis/ gráfico.
Cambio de selec. de cond. de cód. fuente...
Habilitar condiciones de selección
Revisar variables:
Limpiar todas
Aceptar
Cancelar
Abrir...
Guardar como...
Incluir casos
Todo
Específico, seleccionado por:
Por la expresión:
negocio_08>=v7
o nº de caso:
Borrar casos (del conjunto de casos definido en la sección 'Incluir casos')
Por expresión:
o nº de caso:
Por nº de caso:
Ingresar números y/o amplitudes de casos. Ejemplo: 1, 3, 5-12
Por expresión:
Utilice los mismos operadores, funciones y sintaxis de las fórmulas de la hoja de cálculo:
Utilice nombres de variables o v1, v2... v0 es el nº de caso (v0<4 significa casos 1-3)
Ejemplos: (a) v1=0 O edad>18 (b) género='MALE' Y v4<>(v5+v6)
En caso de conflicto, los nombres de las variables tienen prioridad sobre los valores de texto de las variables. Especificar los valores de texto; para ello agregando $, como en 'valor' $.

***EJERCICIO 1.22*.-** Obtén la tabla de frecuencias de la variable **nivel_08**, pero sólo para las empresas que en el año 2008 obtuvieron un volumen de ventas mayor o igual que en el año 2007.

Tabla de frecuencia: nivel_08 (comercio_S1)
Incluir condición: negocio_08>=v7

Categoría	Frecuencia	Acumul. de Frecuencia	%	Acumul. de %
Bajo	11	11	17,46032	17,4603
Medio	34	45	53,96825	71,4286
Alto	18	63	28,57143	100,0000
Faltante	0	63	0,00000	100,0000

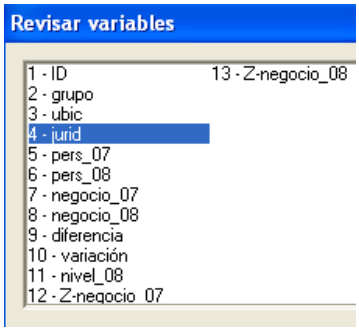
El valor N = 63 nos indica que los cálculos no se han hecho para el archivo completo, sino para los 63 casos que cumplen la condición de selección.

Para desactivar una selección activa, volviendo por tanto a trabajar con el archivo completo, debemos volver a entrar en la selección de casos y deshabilitarla.

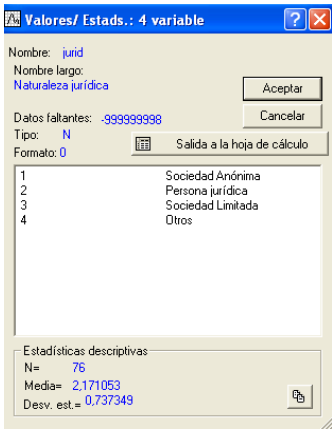
También se puede desactivar desde la barra inferior de la aplicación.

***EJERCICIO 1.23*.-** Realiza un análisis estadístico completo del volumen de negocio en 2008 para las sociedades anónimas.

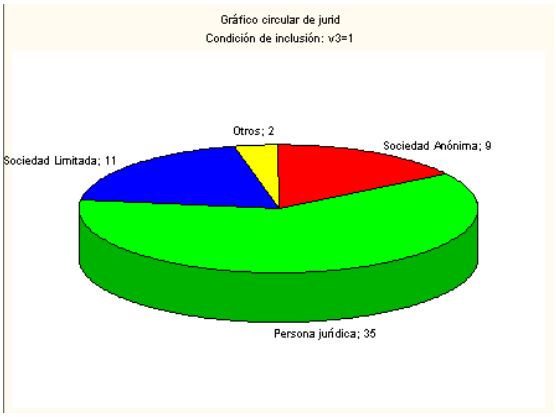
Podemos recordar las variables y, haciendo doble clic, sus valores:



doble clic:

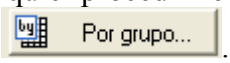


***EJERCICIO 1.24*.-** Realiza un gráfico de sectores que represente la naturaleza jurídica de las empresas situadas fuera de los centros comerciales.

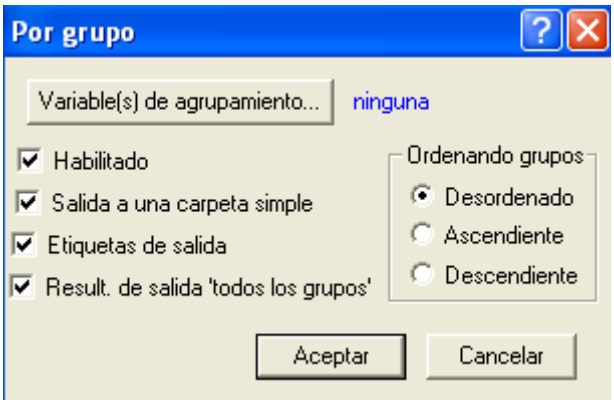


Segmentar (análisis Por Grupos...)

Es habitual, en la mayoría de los análisis estadísticos, que interese realizar, por una parte, un estudio global de todos los datos disponibles, y por otra parte, un estudio desglosado ("segmentado") para distintos subgrupos ("segmentos") existentes en el archivo de datos. Supongamos, por ejemplo, que deseamos estudiar el volumen de negocio en los años 2007 y 2008, no de manera global para todo el archivo, sino desglosando por naturaleza jurídica de las empresas. Deberemos "activar una segmentación por naturaleza jurídica".

En el cuadro de diálogo de cualquier procedimiento estadístico, tenemos un botón que nos permite realizar el análisis 

.



Si ahora indicamos que la variable de agrupamiento es **jurid**, y solicitamos los estadísticos descriptivos por defecto para las variables **negocio_07** y **negocio_08**, obtendremos los resultados para cada valor de la variable **jurid**:

Variable	jurid=Sociedad Anónima Estadísticas descriptivas (comercio_S1)				
	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_07	11	299,9691	75,3500	566,2200	140,6523
negocio_08	11	292,6955	103,5000	502,9300	129,3439

Variable	jurid=Persona jurídica Estadísticas descriptivas (comercio_S1)				
	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_07	45	313,9911	67,52000	667,2400	119,1201
negocio_08	43	318,6084	92,77000	543,0800	108,7520

Variable	jurid=Sociedad Limitada Estadísticas descriptivas (comercio_S1)				
	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_07	16	341,2300	115,7000	606,8500	123,4333
negocio_08	16	320,7356	61,7000	488,0800	129,5818

Variable	jurid=Otros Estadísticas descriptivas (comercio_S1)				
	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_07	4	349,3450	318,0600	380,7900	25,98428
negocio_08	4	372,4650	350,8700	404,4300	24,91264

Para deshabilitar esta opción, basta con volver a entrar en **Por Grupo** y desmarcar la opción de *Habilitar.*

***EJERCICIO 1.25*.-** Realiza un pequeño análisis estadístico en el que podamos analizar el volumen de negocio en 2008 de las empresas, según su ubicación.

	Todos los grupos Estadísticas descriptivas (comercio_S1)				
Variable	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_08	74	318,1276	61,70000	543,0800	113,2829

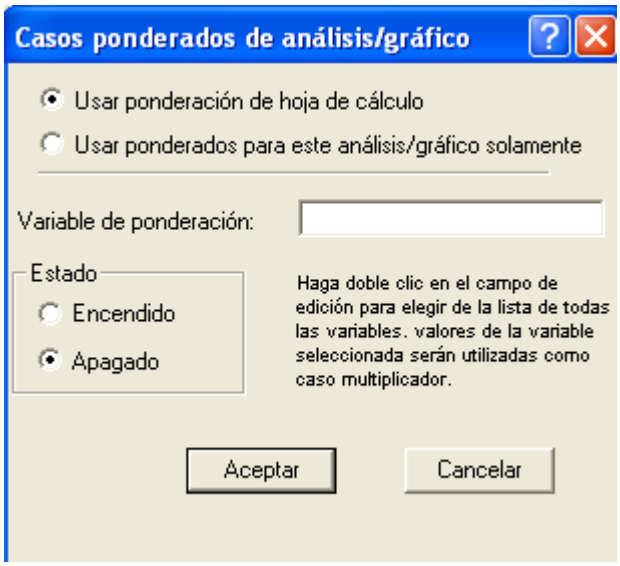
	ubic=Fuera de cualquier centro comercial Estadísticas descriptivas (comercio_S1)				
Variable	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_08	55	307,4078	61,70000	518,5200	114,8192

	ubic=Dentro de un centro comercial Estadísticas descriptivas (comercio_S1)				
Variable	N vál.	Media	Mínimo	Máximo	Desv. est.
negocio_08	19	349,1584	92,77000	543,0800	105,4853

Ponderar

"Ponderar" consiste en asignar pesos (ponderaciones) a los casos de un archivo de datos, con el fin de que unos casos tengan más influencia que otros al realizar un análisis estadístico. Dichos pesos deben estar presentes en una variable del archivo que llamaremos *variable de ponderación*.

En el cuadro de diálogo de cualquier procedimiento estadístico, tenemos un botón que nos permite realizar el análisis ponderado: .


Casos ponderados de análisis/gráfico
Usar ponderación de hoja de cálculo
Usar ponderados para este análisis/gráfico solamente
Variable de ponderación:
Estado
Encendido
Apagado
Haga doble clic en el campo de edición para elegir de la lista de todas las variables. valores de la variable seleccionada serán utilizadas como caso multiplicador.
Aceptar
Cancelar

Veamos un ejemplo de aplicación del comando **Ponderar**. Supongamos que nos hemos encontrado con la siguiente tabla en un informe económico y no tenemos acceso a los datos originales.

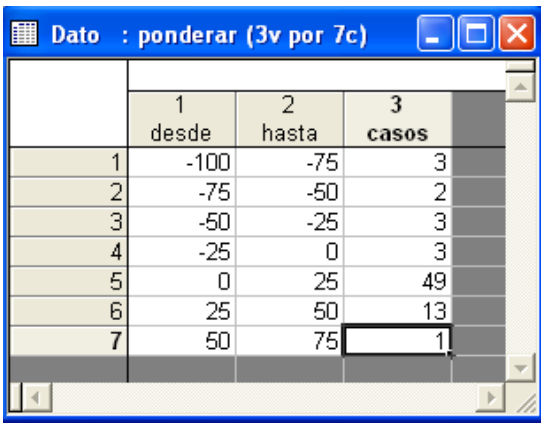
Distribución de la variable que contiene la variación porcentual de 2008 respecto a 2007, agrupada en intervalos

Variación 2007-2008	Empresas
De -100 a -75	3
De -75 a -50	2
De -50 a -25	3
De -25 a 0	3
De 0 a 25	49
De 25 a 50	13
De 50 a 75	1
Total	74

Esta es una manera muy habitual de presentar unos datos estadísticos. Se trata de una distribución de frecuencias con los valores de la variable agrupados en intervalos. Obtenemos una presentación muy compacta pero al precio de perder información si no se tiene acceso a los datos originales.

Por ejemplo, ¿podemos calcular la variación media de estas 74 empresas? La respuesta es NO. Lo que podemos hacer es calcular un valor aproximado de dicha media. El procedimiento a seguir es el siguiente:

1) Introducimos la información de la tabla en un archivo Statistica de la manera siguiente:

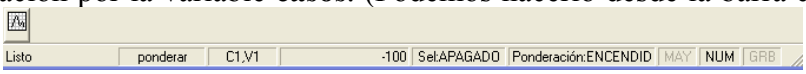


2) Creamos una nueva variable que contenga los "centros" de los intervalos. Estos centros o puntos medios se llaman en Estadística "marcas de clase", y van a actuar como "representantes" de los datos reales a los que no tenemos acceso.

Dato : ponderar* (4v por 7c)

	1 desde	2 hasta	3 casos	4 marca de clase
1	-100	-75	3	-87,5
2	-75	-50	2	-62,5
3	-50	-25	3	-37,5
4	-25	0	3	-12,5
5	0	25	49	12,5
6	25	50	13	37,5
7	50	75	1	62,5

Es la media de estos valores la que nos va a proporcionar una aproximación de la media que nos interesa. El cálculo correcto de tal media debe tener en cuenta que el valor (–87'5) hay que contarlo 3 veces, el valor (–62'5), 2 veces, y así sucesivamente. Lo que tenemos que hacer, entonces, es calcular la media de la variable **marca de clase** usando **casos** como variable de ponderación.

3) Activamos la ponderación por la variable **casos**. (Podemos hacerlo desde la barra de debajo de la aplicación: 

4) Calculamos la media de la variable **marca de clase**.

Variable	Estadísticas descriptivas (ponderar	
	N vál.	Media
marca de clase	74	8,445946

En el archivo de resultados obtenemos finalmente la media buscada. Obsérvese que el valor de N no es 7 (número de casos en el archivo), sino 74 (suma de ponderaciones).

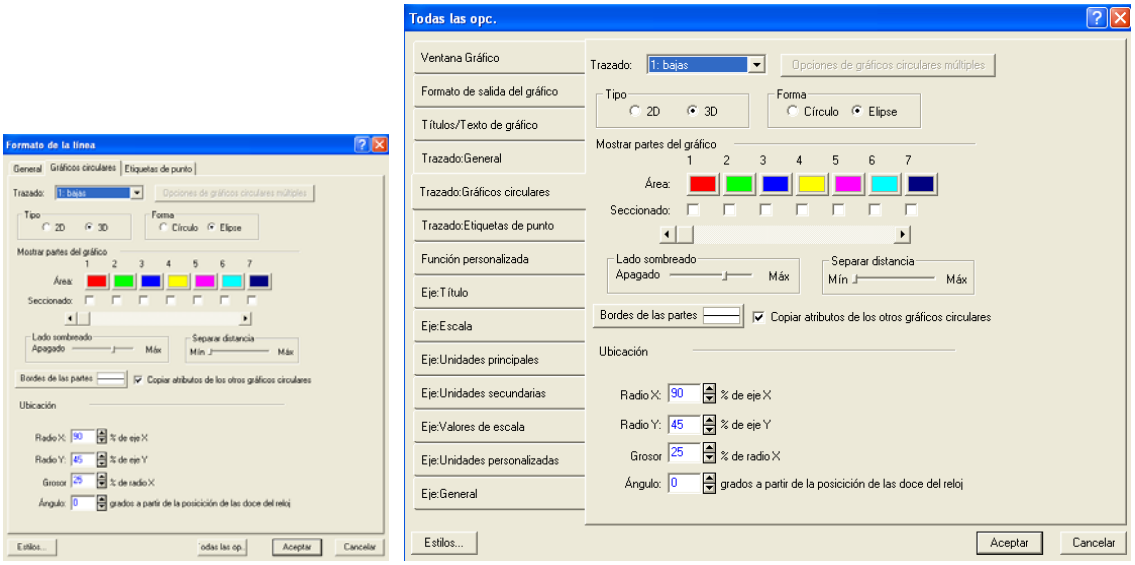
Nuestra aproximación de la media es, pues, 8'4459. Como en realidad sí que tenemos acceso a los datos originales (en el archivo comercio) podemos comprobar que la media "verdadera" es 6'7731.

Variable	Estadísticas descriptivas (comercio_S1	
	N vál.	Media
variación	74	6,773135

5) Desactivamos la ponderación.

1.4 Edición de gráficos.

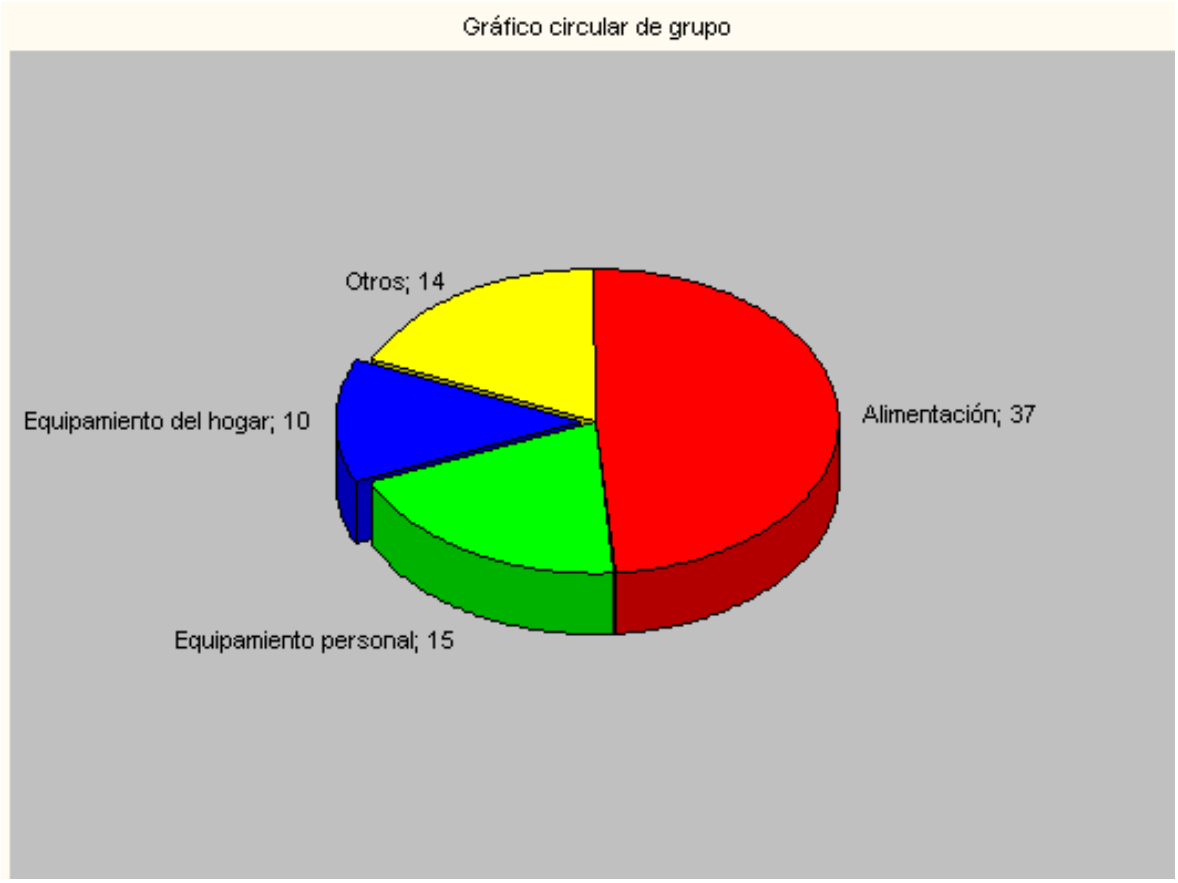
Vamos a realizar un gráfico, editarlo y modificar sus características:
Basta con hacer doble clic en el gráfico para abrir una ventana en la que podemos realizar todas las modificaciones.
Si hacemos doble clic en el dibujo, aparecerá una ventana para modificar las características del gráfico concreto que hayamos editado, mientras que si hacemos doble clic en la zona del gráfico que está "libre", aparecerá una ventana que nos permite modificar cualquier opción de cualquier gráfico.



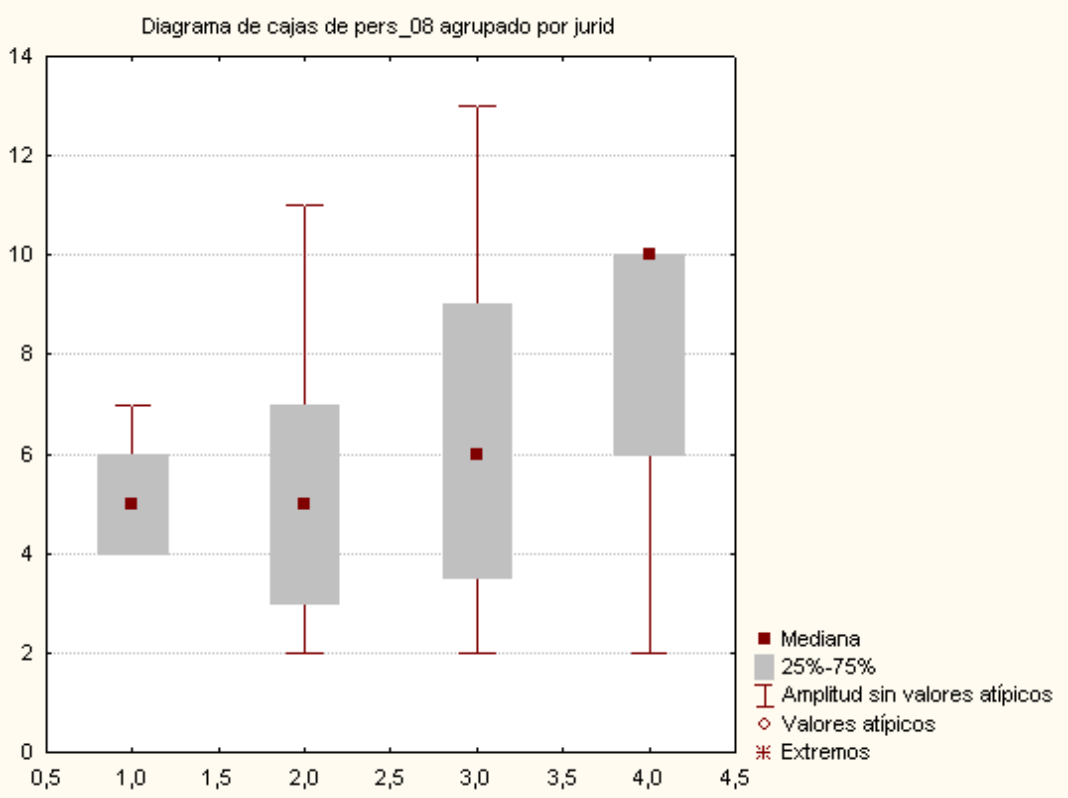
Además, en la ventana de herramientas hay dos pestañas en las que se pueden modificar las opciones por defecto de cualquier gráfico.

Para comprender, y aprender a realizar, las distintas modificaciones lo mejor es practicar, así que lo mejor es que ahora realices los siguientes ejercicios:

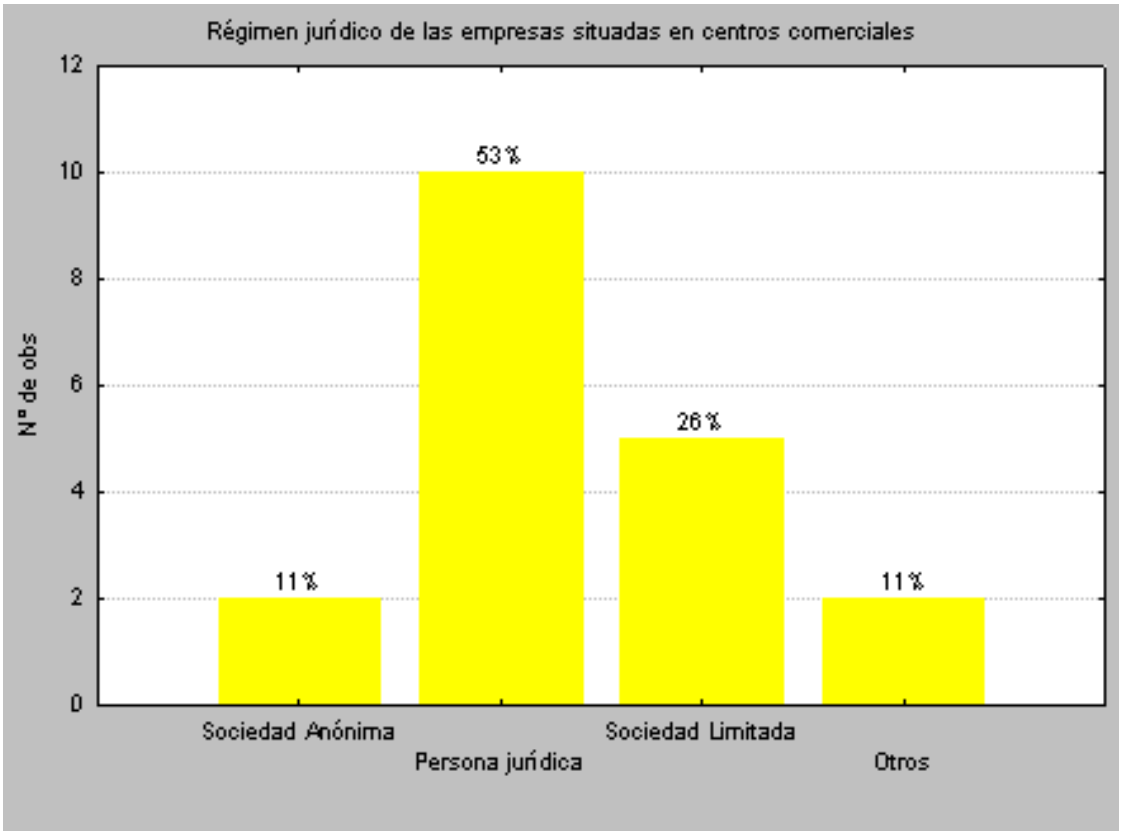
EJERCICIO 1.26.- Reproduce el siguiente gráfico.



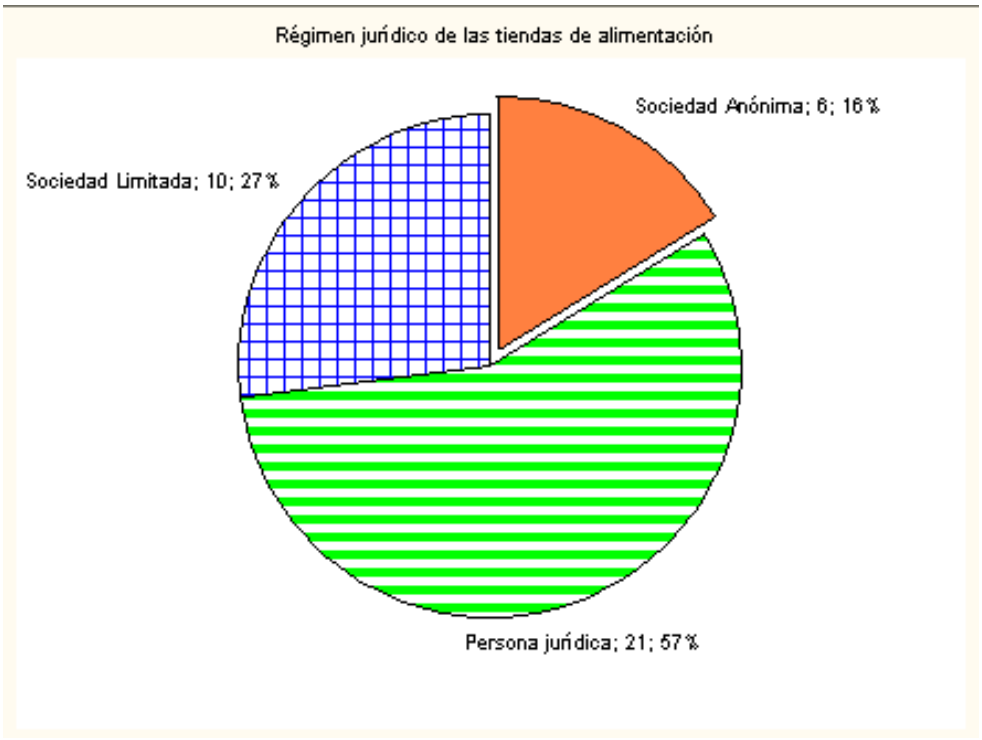
EJERCICIO 1.27.- Reproduce el siguiente gráfico.



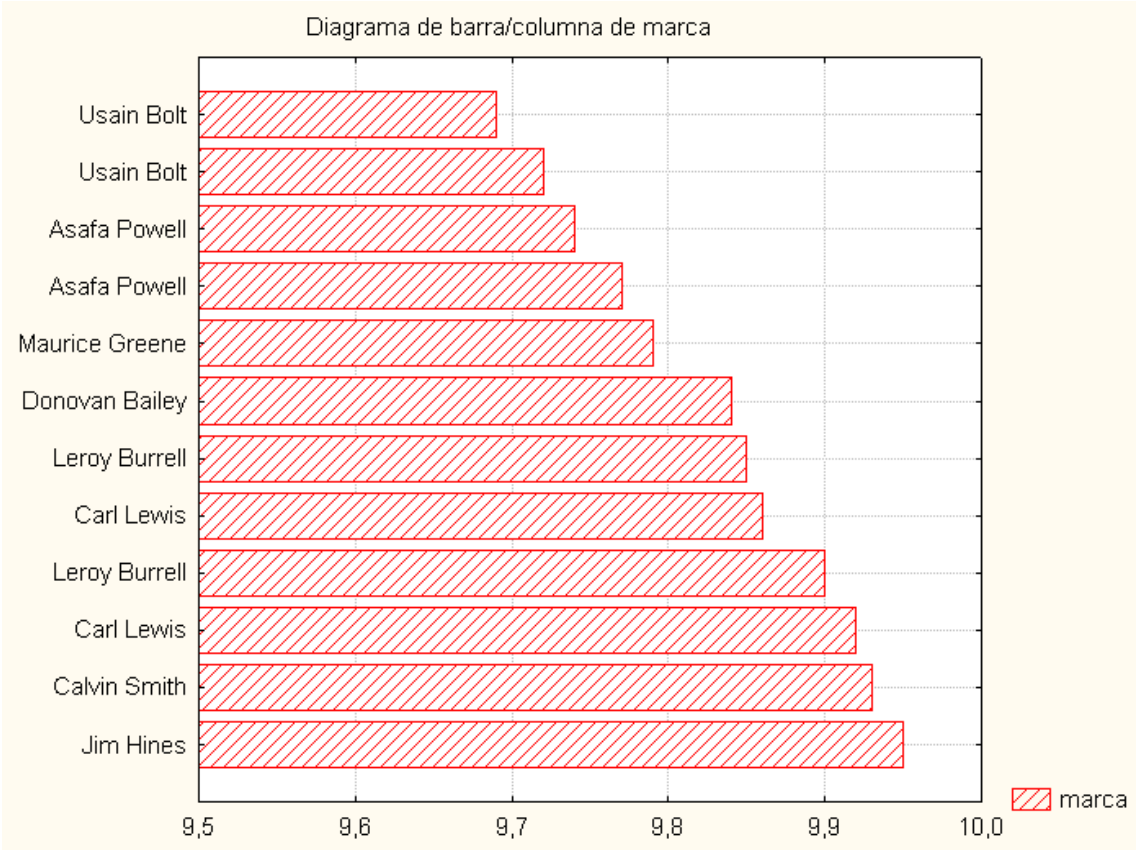
EJERCICIO 1.28.- Reproduce el siguiente gráfico.



EJERCICIO 1.29.- Reproduce el siguiente gráfico.



EJERCICIO 1.30- Reproduce el siguiente gráfico.



Sesión 2. Inferencia Paramétrica

2.1. Inferencia Estadística.

Vamos a enfrentarnos al siguiente problema. Estamos interesados en obtener información de los comercios al por menor de una Comunidad Autónoma durante el año 2007. Estos forman la ***población***, el conjunto de ***individuos*** cuyas características, que llamaremos ***variables,*** queremos estudiar. Por ejemplo, nos centraremos en el estudio del volumen de negocio durante el año 2007, X=NEGOCIO_07.

Como generalmente sucede, no dispondremos de la información de toda la población sino sólo de un subconjunto de la misma que llamaremos ***muestra.*** Así, nuestro objetivo es inferir información para toda la población, a partir de la parcial proporcionada por la muestra. Se puede decir que cualquier estudio de Inferencia Estadística consta de dos pasos:

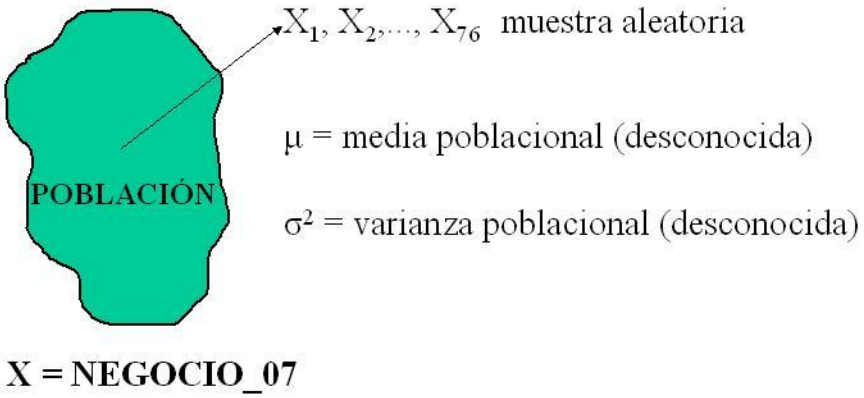
1) Estudio de la muestra aleatoria, utilizando herramientas de la Estadística Descriptiva como las que vimos en la Sesión 1.
2) A partir del paso anterior, generalización para toda la población.

Un primer aspecto al que hemos de prestar atención es que el hecho de poder decir algo fiable acerca de la población a partir de una muestra, depende crucialmente de que dicha muestra sea representativa (respecto a la variable o variables de interés). No nos extenderemos acerca de esta cuestión, que corresponde a la Teoría de Muestras o Métodos de Muestreo. Basta con que digamos que es bastante difícil que una muestra sea representativa si no es aleatoria.

Una rama de la Inferencia Estadística es la Inferencia Paramétrica, que es aquella que se ocupa del estudio de los parámetros de la población (que obviamente se consideran desconocidos): media, varianza, etc. Así, por ejemplo,

- estudiar si una determinada variable X de una población sigue una distribución normal, es un problema de Inferencia no Paramétrica,
- intentar estimar la media, μ, de una variable X de una población, es un problema de Inferencia Paramétrica.

En esta sesión veremos las principales herramientas de la Inferencia Paramétrica. Para ello concretamos nuestro objetivo inicial: estamos interesados en obtener información de la media μ, de la cifra de negocio de los comercios de una Comunidad Autónoma durante el año 2007.


$X_1, X_2, \ldots, X_{76}$ muestra aleatoria
μ = media poblacional (desconocida)
σ^2 = varianza poblacional (desconocida)
POBLACIÓN
X = NEGOCIO_07

El tamaño de la muestra se dice ***tamaño muestral*** y suele denotarse con la letra ***n.*** En nuestro caso ***n*** = 76.

2.2. Intervalos de confianza.

En este apartado nuestro objetivo va a ser estimar el parámetro desconocido μ (media poblacional) de X=NEGOCIO_07. Dicha estimación se basará, obviamente, en la información que nos proporcione la muestra aleatoria. Así, lo primero que vamos a hacer es un estudio descriptivo de la misma.

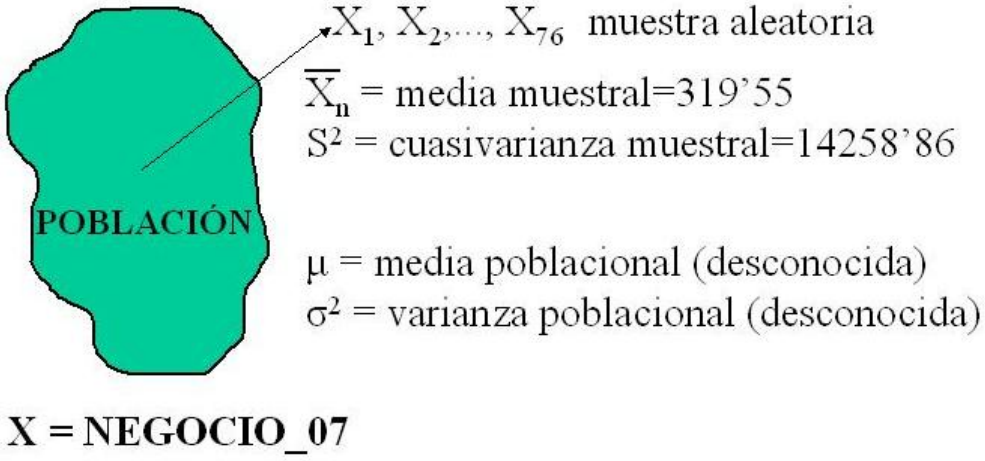
EJERCICIO 2.1.- Calcula la media y la cuasivarianza muestrales de la variable X=NEGOCIO_07.

TABLA 2.1: PARÁMETROS MUESTRALES

$$\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X_n})^2$$

La situación en la que nos encontramos es:


POBLACIÓN
$X_1, X_2, \ldots, X_{76}$ muestra aleatoria
$\overline{X}_n$ = media muestral=319'55
S^2 = cuasivarianza muestral=14258'86
μ = media poblacional (desconocida)
σ^2 = varianza poblacional (desconocida)
X = NEGOCIO_07

Si ahora os pidiese que dieseis una ***estimación puntual*** (es decir, un único valor) para μ y para σ^2, ¿qué diríais?

Existen distintos métodos para dar estimaciones puntuales. Uno de ellos es el método de los momentos, que consiste en estimar los parámetros poblacionales (desconocidos) por los correspondientes muestrales (calculados a partir de la muestra). Además de ser un método muy intuitivo puede demostrarse que, en general, lleva a buenas estimaciones.

Nota importante: una estimación puntual tiene la pega de que no proporciona información acerca del posible error cometido, de la precisión de dicha estimación. Una herramienta que soluciona este problema es la ***estimación por intervalo de confianza*** (algo así como decir: μ, la media poblacional desconocida se encuentra entre 300 y 330). Vamos a hacer un "ejercicio tonto" para introducir algunos conceptos de los intervalos de confianza.

EJERCICIO 2.2.- Proporcionar una estimación por intervalo de la altura del profesor en las siguientes dos condiciones:

SITUACIÓN	**INTERVALO**	**CONFIANZA**	**PRECISIÓN**
Si falláis seréis CASTIGADOS SEVERAMENTE			
Necesito una PRECISIÓN de ±0'5 cms			

Teorema: A partir de una muestra aleatoria $X_1, X_2,...,X_n$ de una variable X que siga una distribución normal, un intervalo de confianza para μ, con nivel de confianza 100x(1-α)%, es

$$IC_{\mu}((1-\alpha)\%) = \overline{X}_n \pm t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}$$

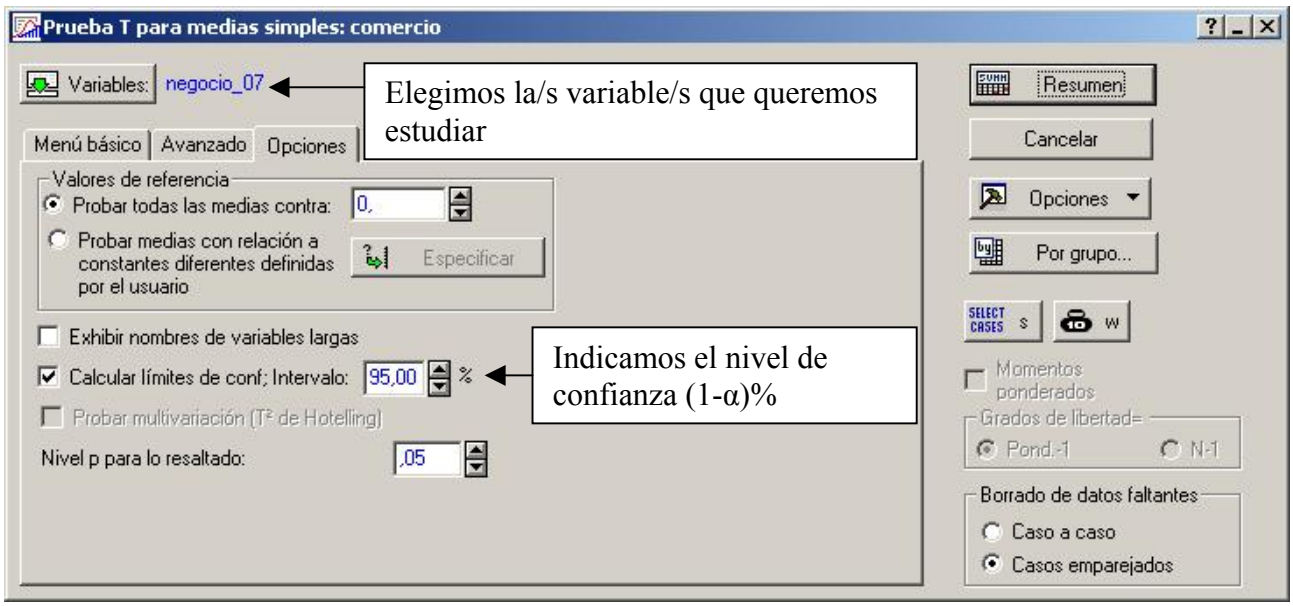
donde $t_{n-1;\alpha/2}$ es el valor que deja a su derecha una probabilidad α/2 en una t_{n-1}

Observación: en la práctica los niveles de confianza habituales son 90%, 95% y 99% (α = 0'1, 0'05, 0'01), según las necesidades de precisión/confianza del parámetro que estemos estimando.

No dedicaremos tiempo a justificar teóricamente el anterior resultado pero sí a notar que la fórmula resulta ser muy intuitiva:

1) El intervalo se construye a partir de la estimación puntual $\overline{X}_n$ (en este caso, de forma simétrica).
2) Los términos que aparecen en la amplitud del intervalo (y que por lo tanto determinan la precisión del mismo) influyen como se supone que deberían hacerlo:
   - cuanta mayor confianza 100x(1-α)% queramos, menor será el valor de α, mayor a su vez el término $t_{n-1;\alpha/2}$ y, por lo tanto, mayor la amplitud del intervalo,
   - cuanto más grande sea el tamaño de la muestra, ***n,*** mayor precisión conseguimos,
   - cuanto menor sea la variabilidad existente en la muestra aleatoria (lo que viene dado por el valor de S), mayor precisión obtendremos.

Para calcular con STATISTICA el intervalo de confianza para la media poblacional desconocida μ acudimos a **Estadísticas/Estadísticas/Tablas básicas,** seleccionamos **Prueba t, muestra simple** y pinchamos en la pestaña **Opciones** donde



y, al hacer clic en ***Resumen*** obtenemos:

Variable	Prueba de medias contra referencia constante (valor) (comercio)									
	Media	Des. est	N	Err.est.	Confianza -95,000%	Confianza +95,000%	Referencia Constante	Valor t	gl	p
negocio_07	319,5568	119,4105	76	13,69732	292,2704	346,8433	0,00	23,32988	75	0,00

EJERCICIO 2.3.- Calcula los intervalos de confianza para la media poblacional μ de X=NEGOCIO_07, con niveles de confianza del 90% y el 99%, notando la relación confianza/precisión.

NIVEL DE CONFIANZA (1-α)%	α	INTERVALO para μ
90%		
99%		

Observaciones:

1) A la hora de interpretar y verbalizar los resultados que hemos obtenido

X = NEGOCIO_07
X es NORMAL

μ = media poblacional (desconocida)

$X_1, X_2, \ldots, X_{76}$ muestra aleatoria

$\overline{X}_n$ = media muestral=319'55

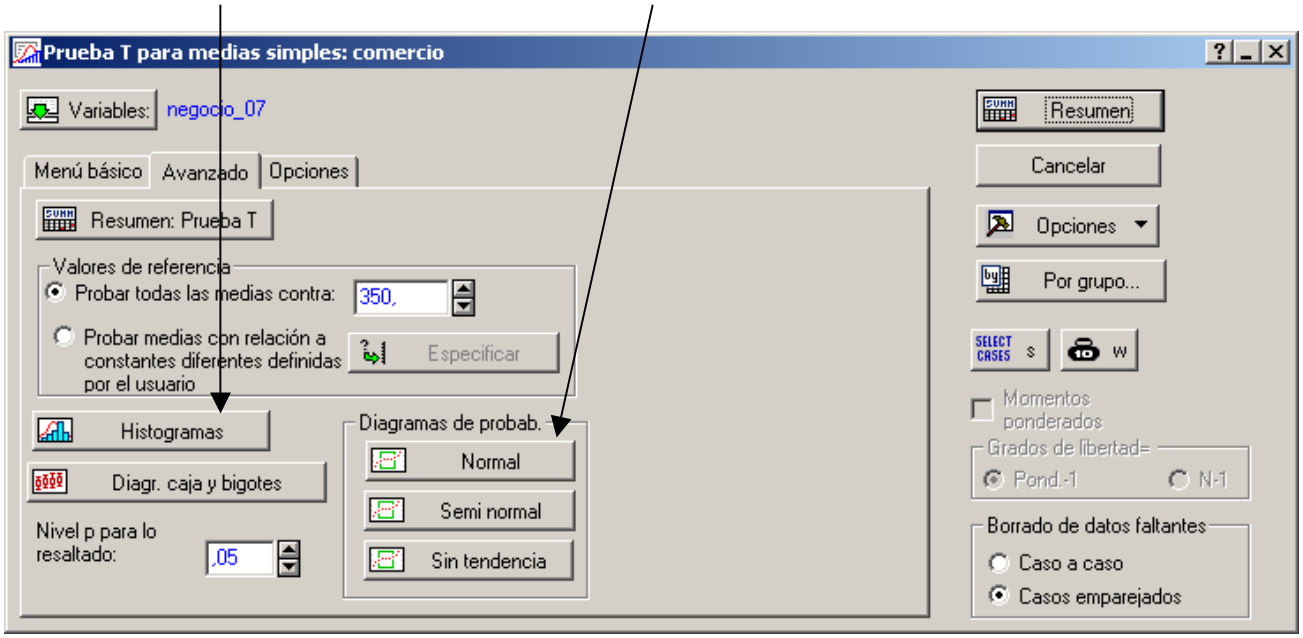
$IC_\mu(95\%)$ = 319'55 ± 27'286 =

= (292'27 , 346'84)

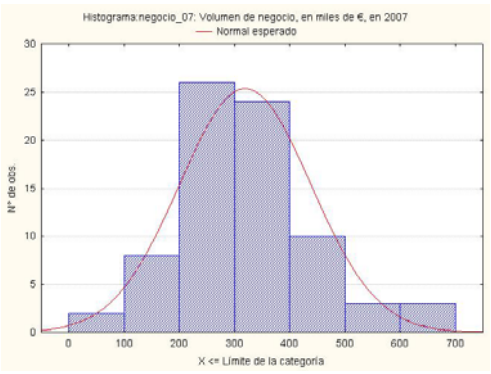
diríamos lo siguiente: a partir de la información proporcionada por una muestra aleatoria $X_1, X_2,...,X_{76}$ de la variable X=NEGOCIO_07 y, respecto a la media poblacional desconocida μ de dicha variable, podemos concluir que,

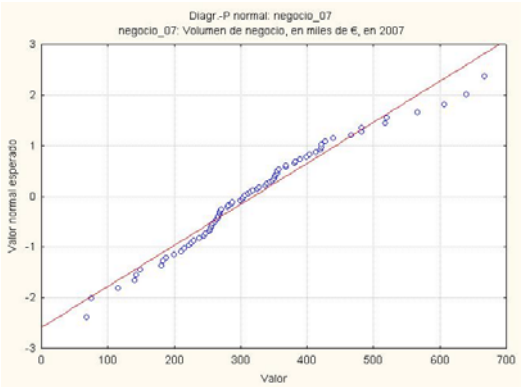
- tenemos una confianza del 95% en que dicho parámetro (desconocido) se encuentre entre 292'27 y 346'84,
- si tomamos el valor de la media muestral $\overline{X}_n = 319'55$ como estimación del parámetro μ (la media poblacional desconocida), "casi seguro", con una confianza del 95%, que nos estamos equivocando en menos de 27'286.

2) Las anteriores conclusiones sólo son válidas <u>si se cumplen las condiciones</u> que permiten construir teóricamente el correspondiente intervalo de confianza, en este caso, la aleatoriedad de la muestra y la normalidad de la población (en lo referido a la variable estudiada). Estas cuestiones serán estudiadas en la Sesión 3. En cualquier caso, hoy anticiparemos aquellas herramientas que el programa nos deja "a tiro". Por ejemplo, si en la ventana anterior pinchamos en la pestaña ***Avanzado*** tenemos la posibilidad de "pegar un vistazo" al cumplimiento de la condición de normalidad representando el correspondiente Histograma con curva normal y/o el Diagrama de normalidad.



obteniéndose:





Y, a la vista del histograma (aparente buen ajuste a la curva normal) y el Diagrama de normalidad (aparente buen ajuste a la recta), podemos decir que la condición de normalidad "tiene pinta" de cumplirse.

3) Los intervalos de confianza son una herramienta muy importante de la Inferencia Paramétrica y se calculan para multitud de parámetros y situaciones. El desarrollo teórico puede ser más o menos complicado, pero su interpretación siempre es la que acabamos de ver.

TABLA 2.2: INTERVALOS DE CONFIANZA

Situación	Estimador puntual	Intervalo de confianza, con confianza $(1-\alpha)\%$
X: $N(\mu,\ \sigma)$ μ y σ desconocidas $X_1, X_2, \ldots, X_n$ m.a.	$\overline{X_n}$	$\overline{X_n} \pm t_{n-1;\ \frac{\alpha}{2}} \frac{S}{\sqrt{n}}$

2.3. Contrastes de hipótesis.

Nos enfrentamos ahora a la siguiente situación

X = NEGOCIO_07

μ= media poblacional desconocida

σ^2= varianza poblacional desconocida

en la que nos piden que realicemos un ***contraste de hipótesis,*** esto es, que tomemos una decisión respecto a:

$$\begin{cases} H_0 \equiv \mu = 350 \\ H_1 \equiv \mu \neq 350 \end{cases}$$

donde H_0 se dice hipótesis nula y H_1 hipótesis alternativa (notar que son complementarias). En este caso 350 es el llamado ***valor de prueba*** que es habitual denotar como μ_0 (en el ejemplo μ_0=350).

Para tomar la decisión contaremos con la información proporcionada por la muestra aleatoria X_1, X_2,...,X_{76}. Además, nos piden que lo hagamos con un ***nivel de significación*** de α=0'05.

Para intentar entender los detalles de esta situación, vamos a establecer un paralelismo con algo que el cine ha hecho muy familiar para todos nosotros. Supongamos que nos llaman para formar parte de un jurado popular y nos piden que tomemos la siguiente decisión:

$$\begin{cases} H_0 \equiv \text{el acusado es INOCENTE} \\ H_1 \equiv \text{el acusado es CULPABLE} \end{cases}$$

¿Cómo tomaremos la decisión?

¿En qué caso rechazaremos H_0?

¿De qué dos formas podríamos equivocarnos? Nuestro sistema legal, ¿cuál de las dos intenta evitar con más énfasis?

¿Os parece que la situación simétrica

$$\begin{cases} H_0 \equiv \text{el acusado es CULPABLE} \\ H_1 \equiv \text{el acusado es INOCENTE} \end{cases}$$

es similar a la anterior?

CONTRASTE DE HIPÓTESIS	JUICIO
Valor de la media poblacional μ (desconocido)	Realidad (desconocida)
Muestra aleatoria	Pruebas
El nivel de significación viene dado por: α=P(RECHAZAR H_0/H_0 CIERTA) siendo habituales en la práctica =0'01, 0'05, 0'1.	Presunción de inocencia: la inocencia sólo se rechaza cuando las pruebas son concluyentes, más allá de cualquier duda razonable.

En general, en cualquier contraste de hipótesis las combinaciones entre la realidad y la conclusión (o sentencia) dan lugar a 4 posibilidades:

CONCLUSIÓN / **REALIDAD**	**ACEPTAR H_0 (declarado INOCENTE)**	**RECHAZAR H_0 (declarado CULPABLE)**
H_0 CIERTA (es INOCENTE)	DECISIÓN CORRECTA	ERROR TIPO I
H_0 FALSA (es CULPABLE)	ERROR TIPO II	DECISIÓN CORRECTA

y ocurre que:

- no es posible construir procedimientos de toma de decisión que minimicen a la vez ambos tipos de error, si intentamos disminuir uno, el otro aumentará; los casos extremos serían:
  - Declarar INOCENTES a todos los acusados: con P(ERROR TIPO I)=0, pero su P(ERROR TIPO II)=1.
  - Declarar CULPABLES a todos los acusados: con P(ERROR TIPO II)=0, pero su P(ERROR TIPO I)=1.
- es habitual imponer que las reglas de decisión tengan una P(ERROR TIPO I)=α (el nivel de significación), que suele tomarse 0'01, 0'05 o 0'1, según las condiciones del problema (especialmente la gravedad de las consecuencias de cometer uno u otro tipo de error), e intentar minimizar el otro error.

Dejemos por un momento el derecho y vamos a centrarnos en nuestro problema:

X = NEGOCIO_07

μ = media poblacional (desconocida)

Contrastar las hipótesis:

$$\begin{cases} H_0 \equiv \mu = 350 \\ H_1 \equiv \mu \neq 350 \end{cases}$$

con nivel de significación α=0'05

$X_1, X_2, \ldots, X_{76}$ muestra aleatoria

$\overline{X}_n$ = media muestral=319'55

y parece intuitivo que nuestro criterio ha de ser:

$$\text{ACEPTAR } H_0 \Leftrightarrow \overline{X}_n \approx \mu_0$$

es decir

$$\text{ACEPTAR } H_0 \Leftrightarrow \overline{X}_n - \mu_0 \approx 0$$

Y "sólo" queda dar rigor a la intuición. Un pequeño desarrollo teórico e, ¡importante!, la suposición de la normalidad de X, permiten traducir $\overline{X}_n - \mu_0 \approx 0$ por

$$-t_{n-1;\alpha/2} \leq \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} \leq t_{n-1;\alpha/2}$$

Observación: Notar que "todo" es intuitivo. Para aceptar la hipótesis nula, exigimos que la media muestral y el valor de prueba sean parecidos, teniendo en cuenta que:

- cuanto mayor sea el tamaño muestral ***n***, más cercanos han de ser ambos valores. Por eso ***n*** aparece multiplicando en la anterior expresión.
- cuanto menor sea la variabilidad de la muestra (esto lo mide ***S***), también más cercanos han de ser ambos valores. Por eso ***S*** aparece dividiendo en la anterior expresión.

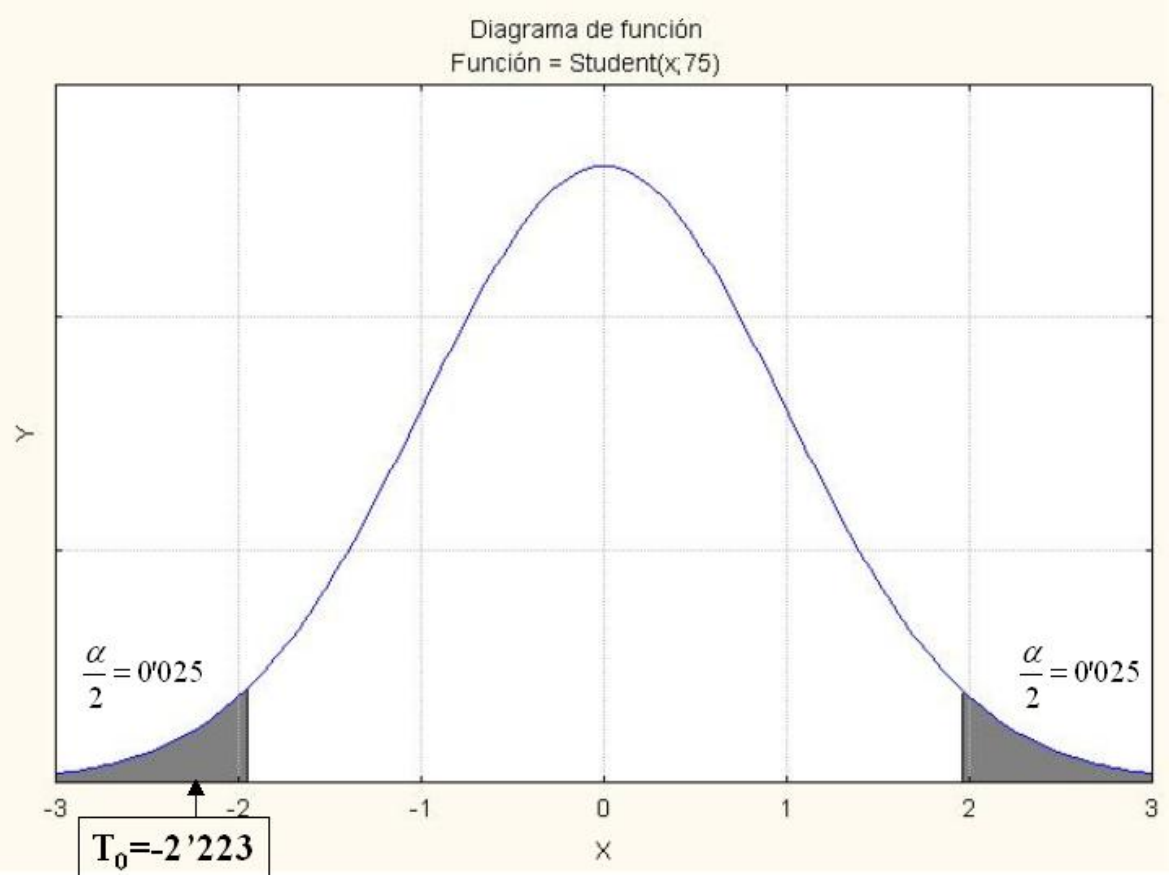
Si hacemos las cuentas en el problema que nos ocupa:

$$t_{n-1;\alpha/2} = t_{75;0'025} = 1'992$$

$$T_0 = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} = \frac{319'55 - 350}{\sqrt{14258'86}/\sqrt{76}} = -2'223$$

Notación: T_0 se dice ***estadístico de prueba*** y $(-\infty, -t_{n-1;\alpha/2}) \cup (t_{n-1;\alpha/2}, +\infty)$ ***región crítica.***

Así, nuestra respuesta al contraste de hipótesis que nos han planteado sería: **a la vista de los datos de la muestra aleatoria, con un nivel de significación α=0'05, RECHAZAMOS la hipótesis nula $H_0 \equiv \mu = 350$.** Gráficamente:



Sin hacer ninguna cuenta adicional, vamos a intentar responder a las siguientes preguntas:

- Si en vez de tomar la decisión con un nivel de significación α=0'05, decidimos hacerlo con α=0'1, ¿cuál sería entonces la conclusión?
- ¿Y si utilizamos α=0'01?

Hay una forma más completa de responder en un contraste de hipótesis, y es calculando el llamado ***p-valor del contraste.*** La idea es sencilla: una vez obtenido T_0 calculamos el área que deja en las dos colas (él y su opuesto) y, comparando con cualquier nivel de significación α dado, podemos responder inmediatamente:

- si *p*-valor > α, entonces ACEPTAMOS H_0, con nivel de significación α,
- si *p*-valor < α, entonces RECHAZAMOS H_0, con nivel de significación α.

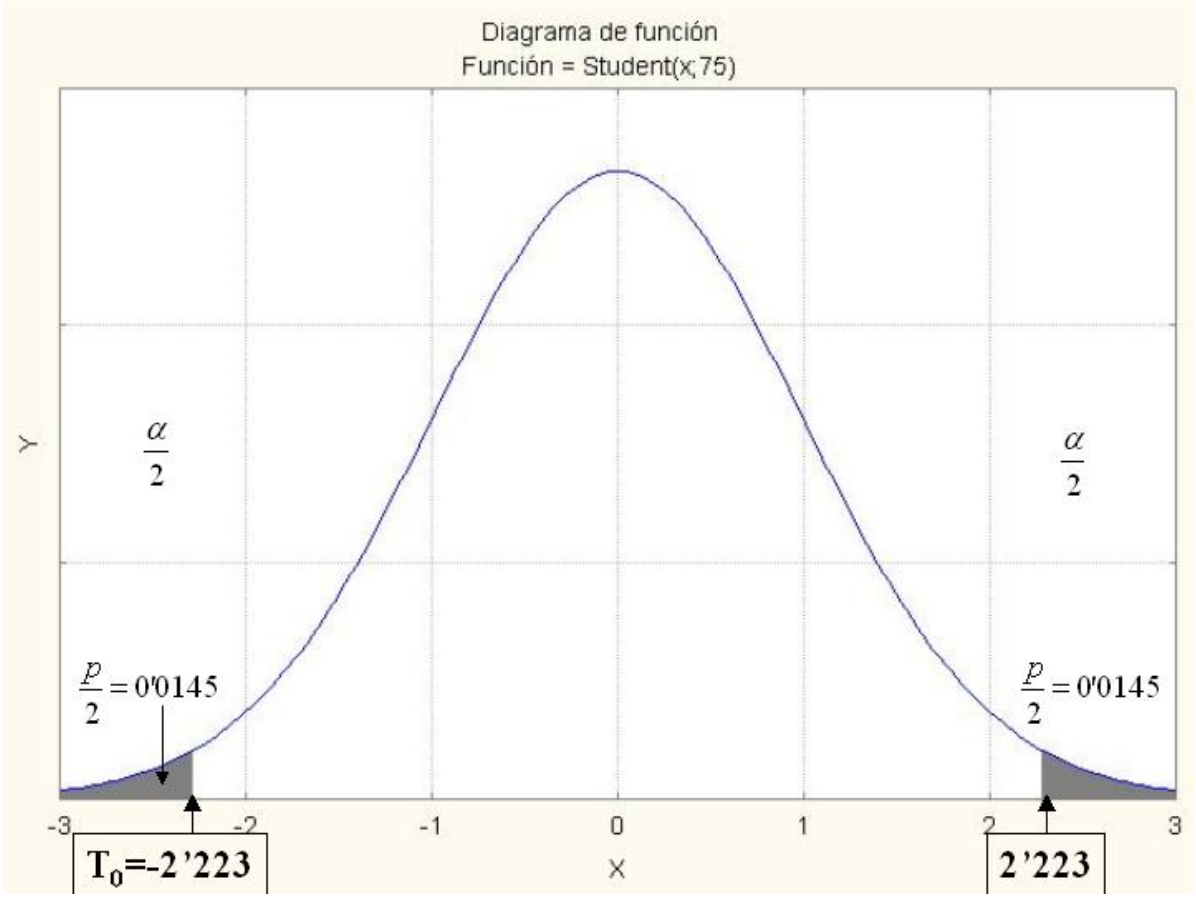
En general, se calcula

$$\frac{p}{2} = P(t_{n-1} < T_0) \quad \text{si } T_0<0 \qquad \text{o,} \qquad \frac{p}{2} = P(t_{n-1} > T_0) \quad \text{si } T_0>0$$

y así, en nuestro ejemplo,

$$\frac{p}{2} = P(t_{n-1} < T_0) = P(t_{75} < -2'223) = 0'0145 \Rightarrow p = 0'029$$

Gráficamente:



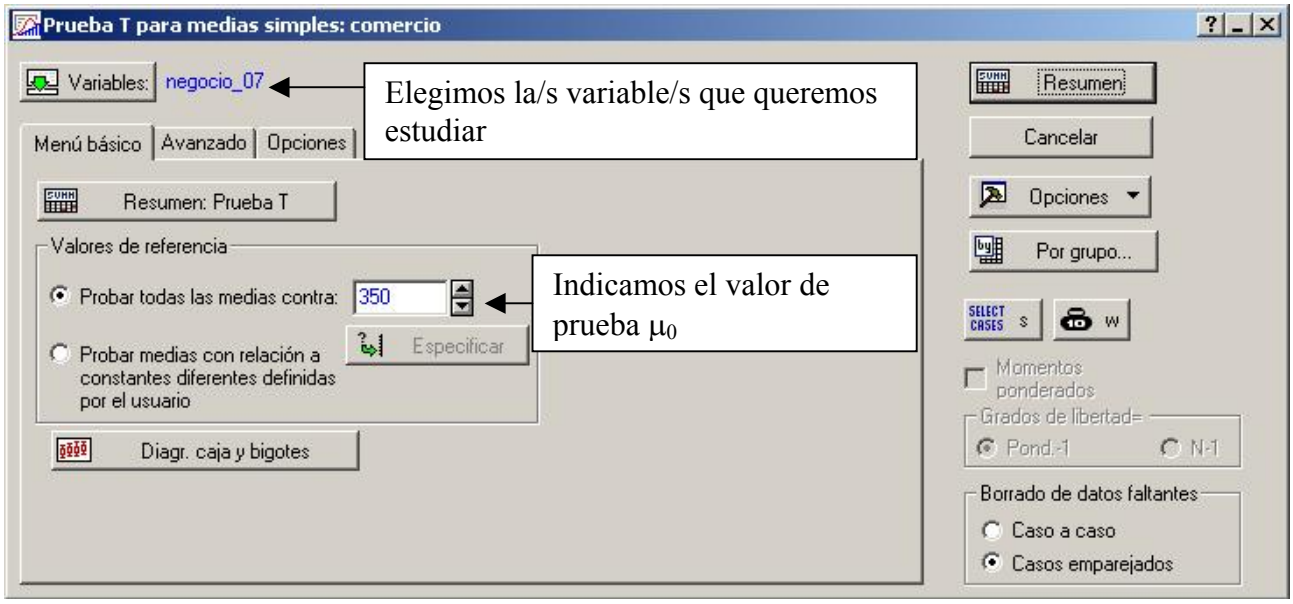
y ahora, podemos responder inmediatamente cuál sería nuestra conclusión para cualquier nivel de significación α.

- para α = 0'05, RECHAZAMOS H_0 porque *p*-valor = 0'029 < α =0'05,
- para α = 0'1, ...
- para α = 0'01, ...
- etc.

Observación:

- un *p*-valor pequeño nos dice que la información proporcionada por la muestra aleatoria proporciona argumentos concluyentes para RECHAZAR H_0. Retomando nuestra afición por el derecho, diríamos que las pruebas dejan clara, más allá de cualquier duda razonable, la culpabilidad del acusado,
- un *p*-valor grande nos dice que la información proporcionada por la muestra aleatoria NO proporciona argumentos suficientes para RECHAZAR H_0. Las pruebas contra el acusado, o son a favor o, aún en contra, dejan dudas razonables.
- un *p*-valor rondando los valores de los niveles de significación α habituales (0'1, 0'05, 0'01) nos obliga a reflexionar (y a definirnos) sobre el margen de error que estamos dispuestos a asumir a la hora de tomar nuestra decisión.

STATISTICA nos hace todas las cuentas anteriores acudiendo a **Estadísticas/Estadísticas/Tablas básicas,** seleccionando **Prueba t, muestra simple**



obteniendo

Variable	Prueba de medias contra referencia constante (valor) (comercio)							
	Media	Des. est	N	Err.est.	Referencia Constante	Valor t	gl	p
negocio_07	319,5568	119,4105	76	13,69732	350,0000	-2,22256	75	0,029259
	$\overline{X}_n$	S			μ_0	T_0	*n-1*	*p-valor*

Notas importantes:

1) En general, existe una relación (muy intuitiva) entre los intervalos de confianza y la respuesta en un contraste de hipótesis. Así, para un nivel de significación α, la respuesta para el contraste de hipótesis

$$\begin{cases} H_0 \equiv \mu = \mu_0 \\ H_1 \equiv \mu \neq \mu_0 \end{cases}$$

será

$$\text{ACEPTAR } H_0 \quad \text{sí y sólo sí} \quad \mu_0 \in IC_\mu(100x(1-\alpha)\%)$$

EJERCICIO 2.4.- Repasad los resultados que hemos ido obteniendo y comprobad el cumplimiento de la anterior propiedad:

100x(1-α)%	IC	¿$350 \in IC_\mu(100x(1-\alpha)\%)$?	α	¿ACEPTAR H_0?
90%				
95%				
99%				

2) Recordad (lo hemos visto al calcular el intervalo de confianza), que haciendo clic en la pestaña ***Avanzado*** podemos estudiar gráficamente la condición de normalidad.

3) Si hemos entendido la mecánica de los anteriores contrastes de hipótesis (que se dicen bilaterales, al poder ser rechazada la hipótesis nula tanto a izquierda como a derecha), no nos sorprenderán los contrastes añadidos en la siguiente tabla resumen (que se dicen unilaterales):

TABLA 2.3: CONTRASTES DE HIPÓTESIS PARA MUESTRAS SIMPLES

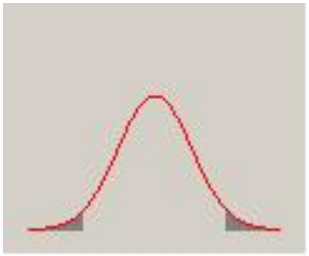
Situación	Hipótesis	Estadístico	R.A.
X: $N(\mu,\ \sigma)$ μ y σ desconocidas $X_1, X_2, \ldots, X_n$ m.a.	$\begin{cases} H_0 \equiv \mu = \mu_0 \\ H_1 \equiv \mu \neq \mu_0 \end{cases}$	$T_0 = \dfrac{\overline{X}_n - \mu_0}{S/\sqrt{n}}$	$\left(-t_{n-1;\ \frac{\alpha}{2}},\ t_{n-1;\ \frac{\alpha}{2}}\right)$
	$\begin{cases} H_0 \equiv \mu \geq \mu_0 \\ H_1 \equiv \mu < \mu_0 \end{cases}$		$\left(-t_{n-1;\ \alpha},\ +\infty\right)$
	$\begin{cases} H_0 \equiv \mu \leq \mu_0 \\ H_1 \equiv \mu > \mu_0 \end{cases}$		$\left(-\infty,\ t_{n-1;\ \alpha}\right)$

Notar en primer lugar que el estadístico que calculamos para tomar la decisión es el mismo en los tres casos y que, intuitivamente, compara el valor de prueba con la media muestral (para que quede más claro utilizaremos los valores del ejercicio que hemos venido haciendo),

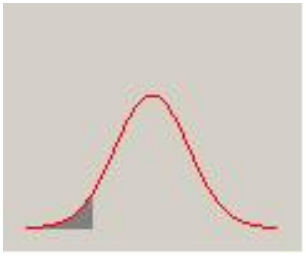
$$T_0 = \frac{\overline{X}_n - \mu_0}{S/\sqrt{n}} = \frac{\overline{X}_n - 350}{\sqrt{14258'86}/\sqrt{76}}$$

lo que cambia es nuestro criterio de decisión (en los siguientes gráficos, el sombreado indica la región crítica, es decir, aquella en la que rechazaríamos H_0, para un nivel de significación α=0'05).

$\begin{cases} H_0 \equiv \mu = 350 \\ H_1 \equiv \mu \neq 350 \end{cases}$ $\begin{cases} H_0 \equiv \mu \geq 350 \\ H_1 \equiv \mu < 350 \end{cases}$ $\begin{cases} H_0 \equiv \mu \leq 350 \\ H_1 \equiv \mu > 350 \end{cases}$



Rechazamos H_0 cuando $\overline{X}_n$ esté "lejos" μ_0=350, es decir, cuando T_0 sea, o muy positivo o muy negativo.



Rechazamos H_0 cuando $\overline{X}_n$ sea muy "pequeño" comparado con μ_0=350, es decir, cuando T_0 sea muy negativo.

Rechazamos H_0 cuando $\overline{X}_n$ sea muy "grande" comparado con μ_0=350, es decir, cuando T_0 sea muy negativo.

y, en cada uno de los casos, el p-valor se calcularía:

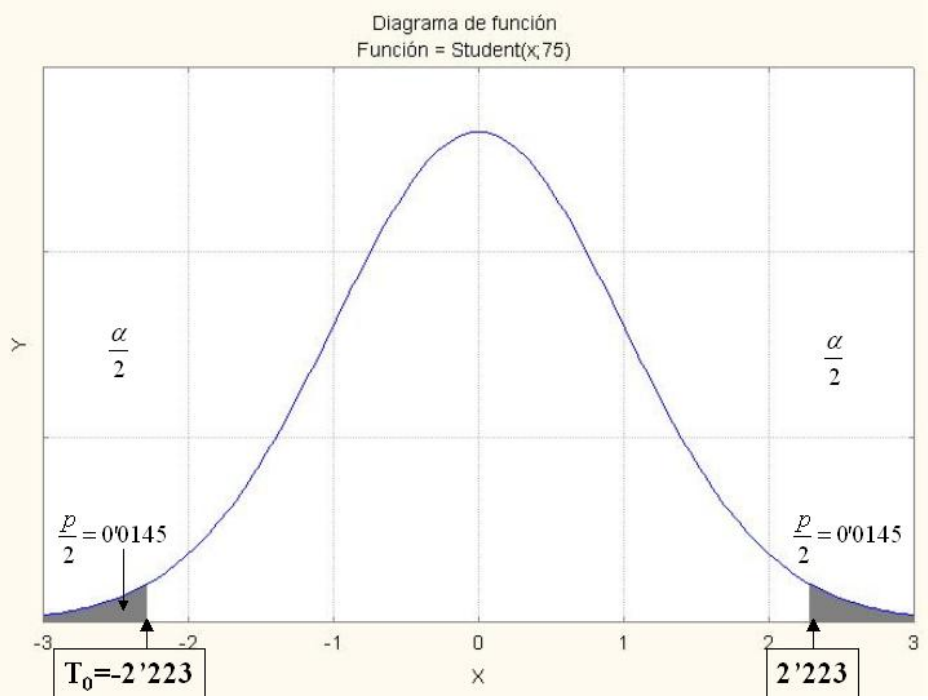
$\dfrac{p}{2} = P(t_{n-1} < T_0)$ si T_0<0

o,

$p = P(t_{n-1} < T_0)$ $\qquad$ $p = P(t_{n-1} > T_0)$

$\dfrac{p}{2} = P(t_{n-1} > T_0)$ si T_0>0

¿Nos hace STATISTICA las cuentas? Pues en realidad NO, el programa sólo nos calcula el p-valor del contraste bilateral (lo hemos hecho antes a partir de una media muestral: $\overline{X}_n = 319'55$, –ver el siguiente gráfico- y hemos obtenido p=0’029, es decir, $\frac{p}{2} = 0'0145$).



Pero eso no es un gran problema: ¿os atrevéis a completar el siguiente cuadro?

$\begin{cases} H_0 \equiv \mu = 350 \\ H_1 \equiv \mu \neq 350 \end{cases}$	$\begin{cases} H_0 \equiv \mu \geq 350 \\ H_1 \equiv \mu < 350 \end{cases}$	$\begin{cases} H_0 \equiv \mu \leq 350 \\ H_1 \equiv \mu > 350 \end{cases}$
p-valor=0’029	p-valor=	p-valor=
Decisión para α=0’05	Decisión para α=0’05	Decisión para α=0’05

En general, a partir del p-valor calculado para un contraste bilateral (llamémosle P_B, que es el que nos calcula STATISTICA), el p-valor para los contrastes unilaterales se obtendría:

$\begin{cases} H_0 \equiv \mu \geq \mu_0 \\ H_1 \equiv \mu < \mu_0 \end{cases}$	p= $P_B/2$ cuando $T_0 < 0$ p=1- $P_B/2$ cuando $T_0 > 0$ (evidencia a favor de H_0).
$\begin{cases} H_0 \equiv \mu \leq \mu_0 \\ H_1 \equiv \mu > \mu_0 \end{cases}$	p= $P_B/2$ cuando $T_0 > 0$ p=1- $P_B/2$ cuando $T_0 < 0$ (evidencia a favor de H_0).

2.4. Contrastes paramétricos de dos poblaciones.

Muestras independientes

Nos enfrentamos al siguiente problema:

X = NEGOCIO_07 (FUERA de los Centros Comerciales)
Y = NEGOCIO_07 (DENTRO de los Centros Comerciales)

μ_X= media desconocida de la población X
σ_X^2= varianza desconocida de la población X

μ_Y= media desconocida de la población Y
σ_Y^2= varianza desconocida de la población Y

y queremos contrastar las hipótesis:

$$\begin{cases} H_0 \equiv \mu_X = \mu_Y \\ H_1 \equiv \mu_X \neq \mu_Y \end{cases}$$

y para tomar la decisión contamos con la información proporcionada por dos muestras aleatorias de tamaños ***n***=57 y ***m***=19

$X_1, X_2, \ldots, X_{57}$ muestra aleatoria de X

$\overline{X}_n = 320'38$

$Y_1, Y_2, \ldots, Y_{19}$ muestra aleatoria de Y

$\overline{Y}_m = 317'06$

y parece intuitivamente razonable tomar la decisión comparando los valores de las medias muestrales y:

$$\text{ACEPTAR } H_0 \Leftrightarrow \overline{X}_n \approx \overline{Y}_m \quad \text{es decir} \quad \text{ACEPTAR } H_0 \Leftrightarrow \overline{X}_n - \overline{Y}_m \approx 0$$

Un pequeño desarrollo teórico (suponiendo que se cumplen ciertas hipótesis: aleatoriedad de las muestras, independencia de X e Y, normalidad de X e Y, y homoscedasticidad –igualdad de las varianzas poblacionales-), nos lleva al siguiente estadístico de prueba:

$$T_0 = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_X^{\,2}+(m-1)S_Y^{\,2}}{n+m-2}}}$$

y a la región crítica para un nivel de significación α, $(-\infty, -t_{n+m-2;\alpha/2}) \cup (t_{n+m-2;\alpha/2}, +\infty)$ o, siempre la mejor opción, calcular el *p*-valor del contraste como:

$$\frac{p}{2} = P(t_{n+m-2} < T_0) \quad \text{si } T_0<0 \qquad \text{o,} \qquad \frac{p}{2} = P(t_{n+m-2} > T_0) \quad \text{si } T_0>0$$

En nuestro caso:

$$T_0 = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1).S_X^2+(m-1).S_Y^2}{n+m-2}}} = \frac{320'38-317'06}{31'84} = 0'104$$

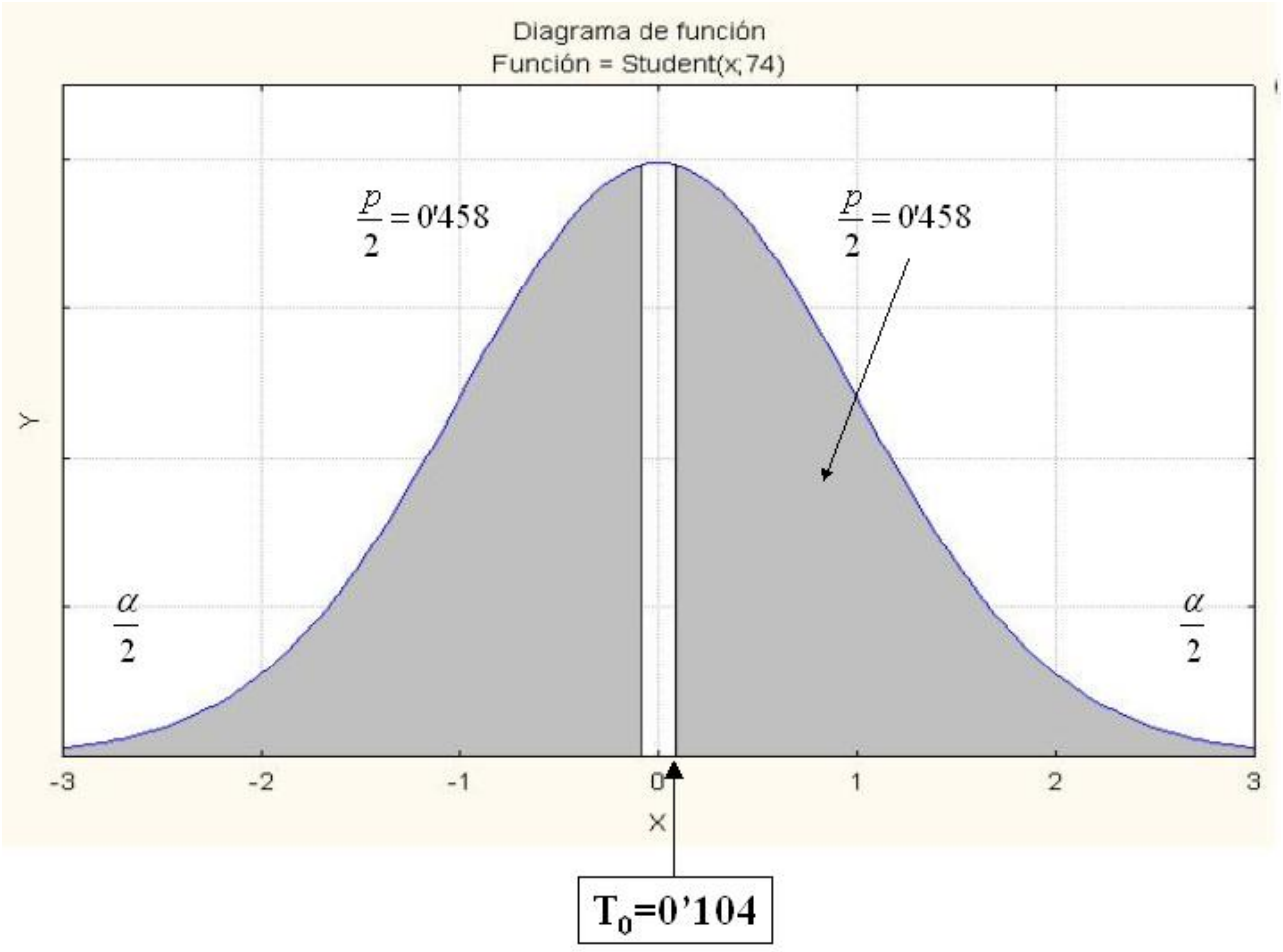
$$\frac{p}{2} = \mathrm{P}(t_{n+m-2} > T_0) = \mathrm{P}(t_{74} > 0'104) = 0'458$$

es decir,

***p*-valor = 0'917 > α = 0'05**

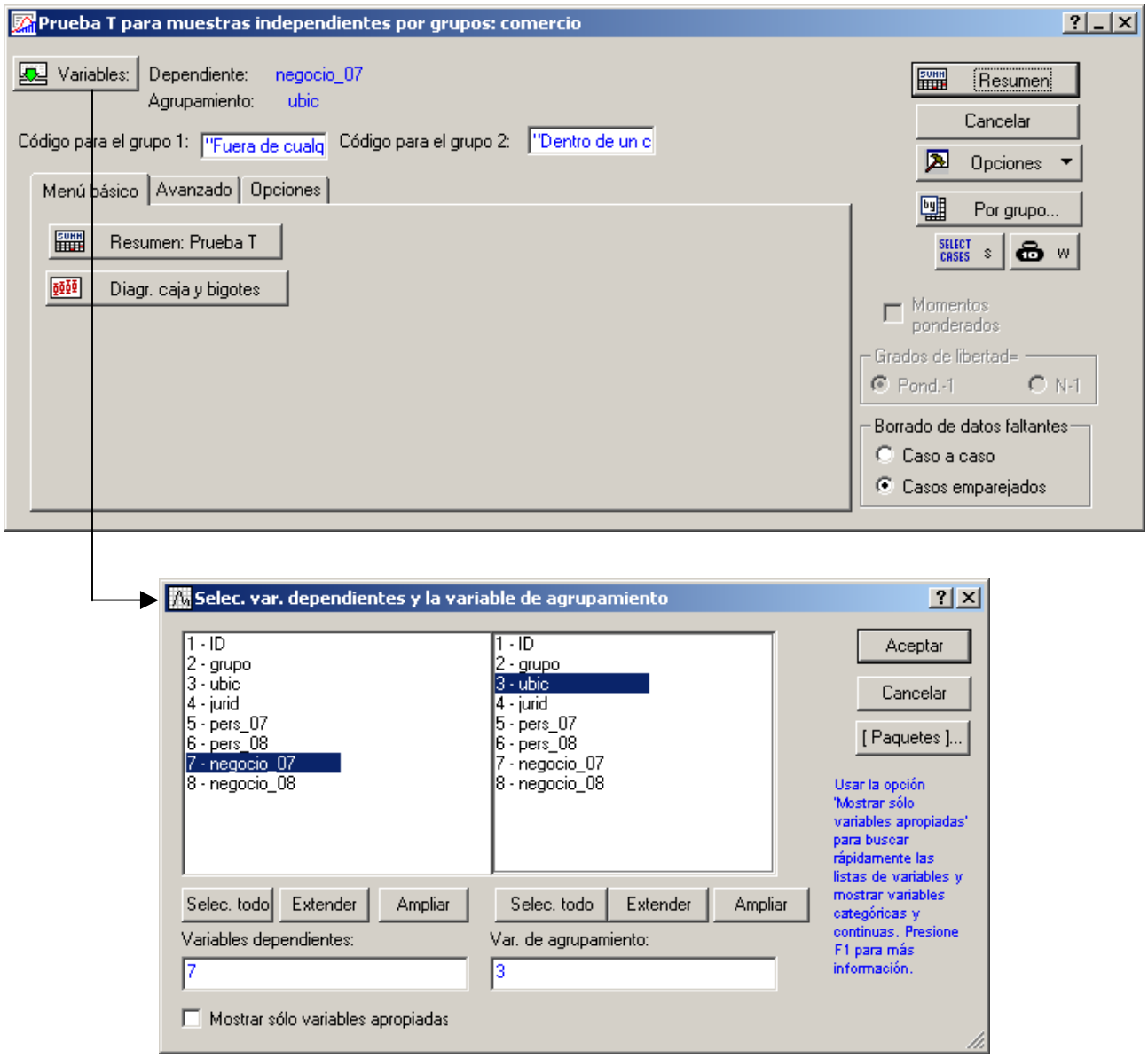
ACEPTAMOS H_0

Gráficamente:



Nota: en este caso el *p*-valor es "enorme" (sobre todo si lo comparamos con los niveles de significación habituales, 0'01, 0'05 o 0'1), lo que, como hemos comentado, las muestras apoyan la veracidad de la hipótesis nula.

STATISTICA nos hace las cuentas en **Estadísticas/Estadísticas/Tablas básicas,** seleccionando **Prueba t, independiente, por grupos** (ya que tenemos la información almacenada en dos grupos de una misma variable. Si la tuviésemos en dos variables distintas elegiríamos **Prueba t, independiente, por variables**).



Y pulsando en ***Resumen*** obtenemos:

Variable	Media Fuera de cualquier centro comercial	Media Dentro de un centro comercial	Valor t	gl	p	N vál. Fuera de cualquier centro comercial	N vál. Dentro de un centro comercial	Desv. est. Fuera de cualquier centro comercial	Desv. est. Dentro de un centro comercial	Razón-F Varianzas	p Varianzas
negocio_07	320,3872	317,0658	0,104305	74	0,917210	57	19	123,1613	110,5064	1,242147	0,627655

$\overline{X}_n$ $\overline{Y}_m$ *p-valor* S_X S_Y

$$T_0 = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{1}{n}+\frac{1}{m}}\sqrt{\frac{(n-1)S_X^{\,2}+(m-1)S_Y^{\,2}}{n+m-2}}}$$

Contraste de homoscedasticidad

Observaciones:

- por defecto el programa realiza un contraste para comprobar la condición de homoscedasticidad.

$$\begin{cases} H_0 \equiv \sigma_X^2 = \sigma_Y^2 \\ H_1 \equiv \sigma_X^2 \neq \sigma_Y^2 \end{cases}$$

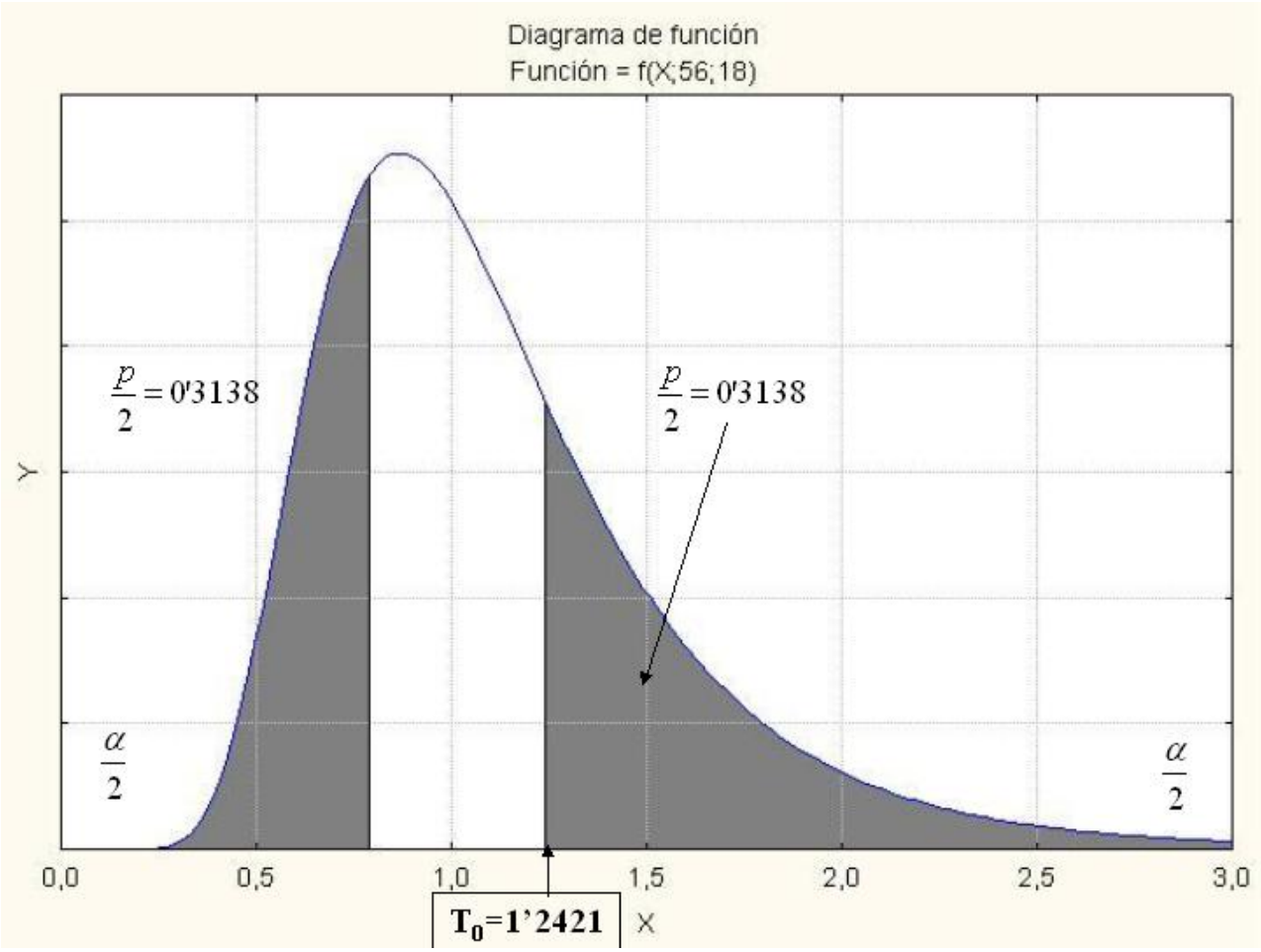
que utiliza como estadístico de contraste $T_0 = \frac{S_X{}^2}{S_Y{}^2}$ Intuitivamente, ¿cuándo aceptaremos H_0?

El correspondiente desarrollo teórico (suponiendo la aleatoriedad de las muestras y la independencia y normalidad de X e Y), nos llevan a la región crítica para un nivel de significación α, $(0, F_{n-1,m-1;1-\alpha/2}) \cup (F_{n-1,m-1;\alpha/2}, +\infty)$ o, calcular el *p*-valor del contraste como:

$$\frac{p}{2} = P(F_{n-1,m-1} < T_0) \quad \text{si } T_0 \text{ está en la cola izquierda} \quad \text{o,}$$

$$\frac{p}{2} = P(F_{n-1,m-1} > T_0) \quad \text{si } T_0 \text{ está en la cola derecha}$$

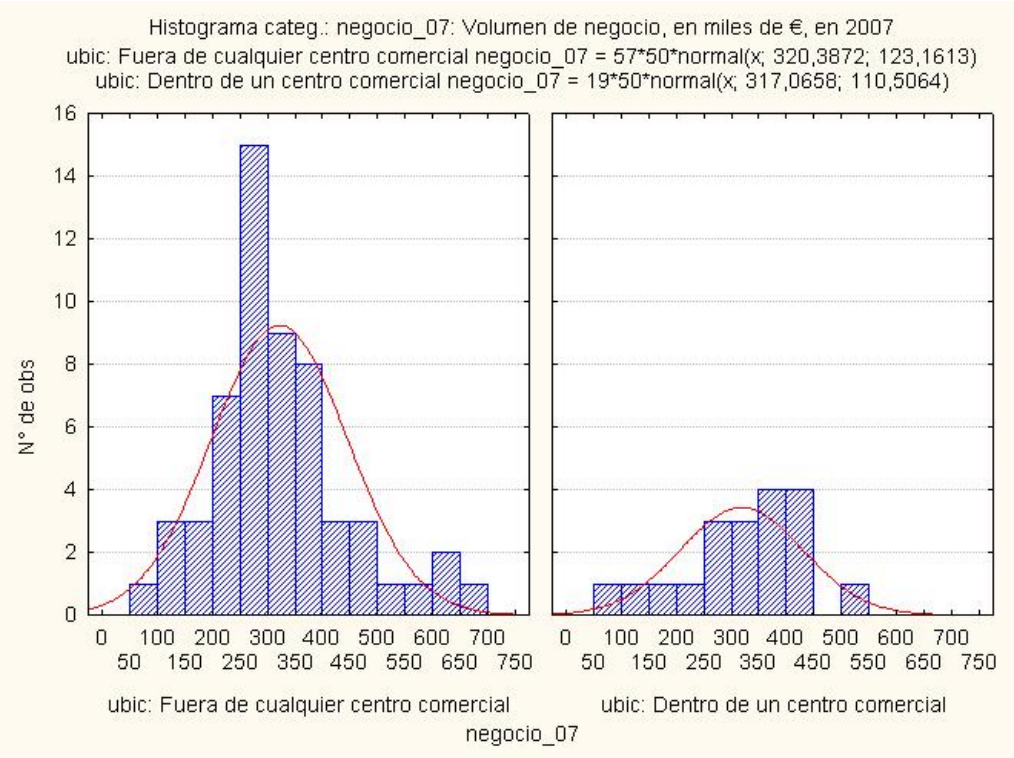
En nuestro caso, hemos obtenido:

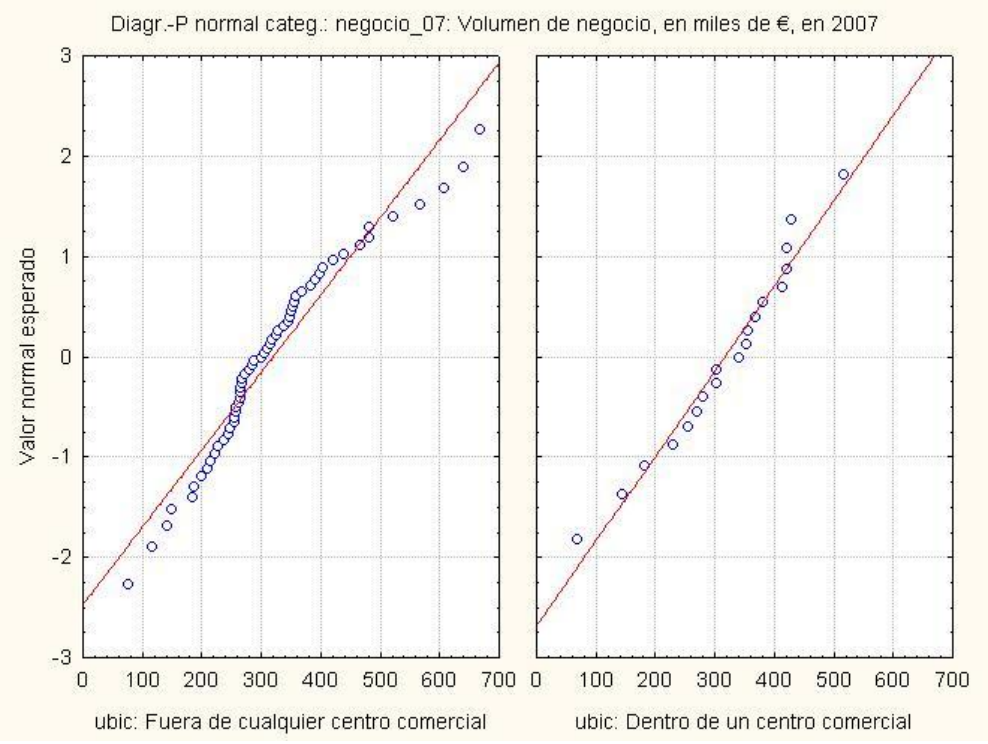


Y así, ¿aceptamos o rechazamos la hipótesis de homoscedasticidad?

- el contraste de homoscedasticidad que acabamos de ver es muy sensible al cumplimiento de la condición de normalidad. El programa ofrece como alternativas, en la pestaña ***Opciones,*** la ***Prueba de Levene*** y la de ***Brown y Forshyte,*** que estudiaremos en la próxima sesión.

- en la pestaña ***Avanzado*** podemos representar gráficos que nos permiten hacernos una idea del cumplimiento de la condición de normalidad (eligiendo, por ejemplo, ***Histogramas categorizados*** y/o ***Diagr. categorizados normales***). Obtendríamos:


Histograma categ.: negocio_07: Volumen de negocio, en miles de €, en 2007
ubic: Fuera de cualquier centro comercial negocio_07 = 57*50*normal(x; 320,3872; 123,1613)
ubic: Dentro de un centro comercial negocio_07 = 19*50*normal(x; 317,0658; 110,5064)
N° de obs
ubic: Fuera de cualquier centro comercial
ubic: Dentro de un centro comercial
negocio_07


Diagr.-P normal categ.: negocio_07: Volumen de negocio, en miles de €, en 2007
Valor normal esperado
ubic: Fuera de cualquier centro comercial
ubic: Dentro de un centro comercial

¿Cómo "pinta" la cosa?

Muestras dependientes

Vamos a intentar resolver ahora una situación ¿similar?

X = NEGOCIO_08 (Volumen de negocio en 2008)
Y = NEGOCIO_07 (Volumen de negocio en 2007)

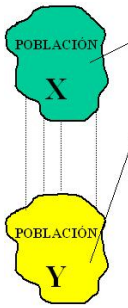
μ_X= media desconocida de la población X
σ_X^2= varianza desconocida de la población X

μ_Y= media desconocida de la población Y
σ_Y^2= varianza desconocida de la población Y

y queremos contrastar las hipótesis:

$$\begin{cases} H_0 \equiv \mu_X = \mu_Y \\ H_1 \equiv \mu_X \neq \mu_Y \end{cases}$$

Para tomar la decisión contamos con la información proporcionada por una muestra aleatoria de ***n***=74 parejas de observaciones de X e Y.



$X_1, X_2, ..., X_{74}$ muestra aleatoria de X
$Y_1, Y_2, ..., Y_{74}$ muestra aleatoria de Y

$\overline{X}_n$= **media muestral de X** = 318'12

$\overline{Y}_n$= **media muestral de Y** = 319'85

S^2_{X-Y} = **varianza muestral de X-Y** = 18400'68

EJERCICIO 2.5.- Comprueba que la varianza muestral de X-Y = NEGOCIO_08-NEGOCIO_07, es la indicada. (Recuerda que dicha variable la creamos en la Sesión 1 con el nombre DIFERENCIA).

Esa es la novedad de esta situación respecto a la que hemos estudiado en el apartado anterior: las muestras no son independientes ya que provienen de los mismos 74 comercios en dos momentos distintos. Habitualmente ésta es una cuestión de diseño experimental:

- en las **muestras independientes** se dispone de dos grupos seleccionados aleatoriamente y no necesariamente del mismo tamaño. Por ejemplo, cuando para estudiar la eficacia de un medicamento se seleccionan aleatoriamente dos grupos de enfermos: a un grupo se le aplica el tratamiento y al otro no.
- en las **muestras dependientes** (o **pareadas**) se diseña el experimento buscando una subdivisión en dos grupos de igual tamaño, de manera que puedan formarse parejas con un individuo de cada grupo, de forma que sean "iguales o muy parecidos". La idea es "igualar" la influencia de posibles variables no controladas en el estudio. Por ejemplo cuando se estudia la eficacia de un medicamento con un grupo de gemelos enfermos. A uno se le aplica el tratamiento y al otro no. En este caso, tenemos dos grupos: gemelos a los que se aplica el tratamiento y gemelos a los que no. La ventaja es el control sobre las variables que pueden influir en el experimento. La pega es que son diseños mucho más costosos y difíciles de conseguir.

De nuevo, un pequeño desarrollo teórico y la suposición de ciertas condiciones (aleatoriedad de las muestras y normalidad de X-Y), nos permiten construir una regla de decisión a partir de la idea intuitiva

ACEPTAR $H_0 \Leftrightarrow \overline{X}_n \approx \overline{Y}_m$ es decir ACEPTAR $H_0 \Leftrightarrow \overline{X}_n - \overline{Y}_m \approx 0$

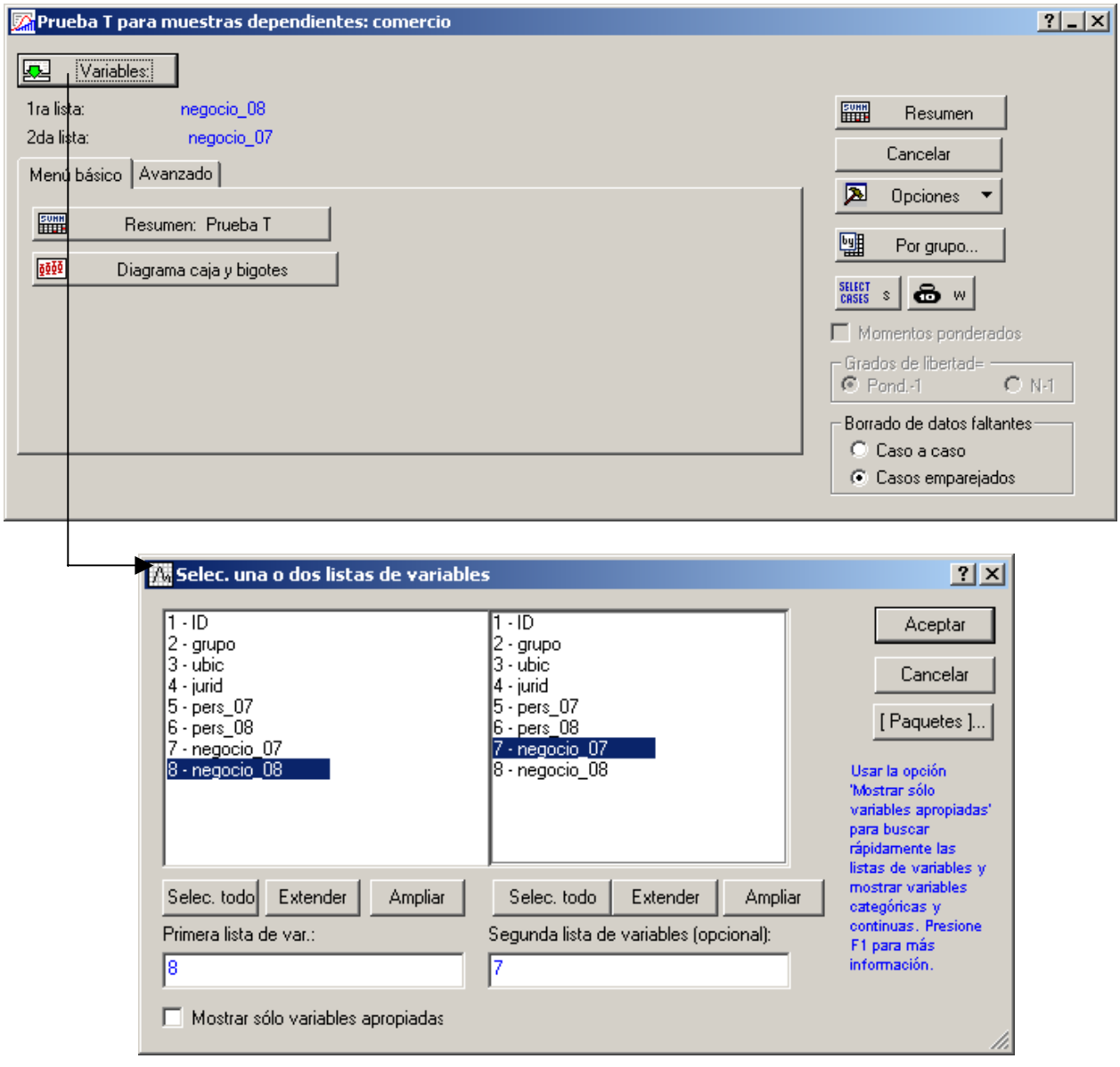
y nos lleva al estadístico de prueba:

$$T_0 = \frac{\overline{X}_n - \overline{Y}_m}{S_{X-Y}/\sqrt{n}}$$

y a la región crítica para un nivel de significación α, $(-\infty, -t_{n-1;\alpha/2}) \cup (t_{n-1;\alpha/2}, +\infty)$ o, siempre la mejor opción, calcular el *p*-valor del contraste como:

$$\frac{p}{2} = P(t_{n-1} < T_0) \quad \text{si } T_0<0 \qquad \text{o,} \qquad \frac{p}{2} = P(t_{n-1} > T_0) \quad \text{si } T_0>0$$

STATISTICA nos hace las cuentas en **Estadísticas/Estadísticas/Tablas básicas,** seleccionando **Prueba t, muestras dependientes**



obteniéndose,

Variable	Prueba t para muestras dependientes (comercio) Diferencias marcadas son significantes con p < ,05000							
	Media	Des. est	N	Dif.	Des. est Dif.	t	gl	p
negocio_08	318,1276	113,2829						
negocio_07	319,8534	120,8128	74	-1,72581	135,6491	-0,109444	73	0,913151

$\overline{X}_n - \overline{Y}_m$ S_{X-Y} *p-valor*

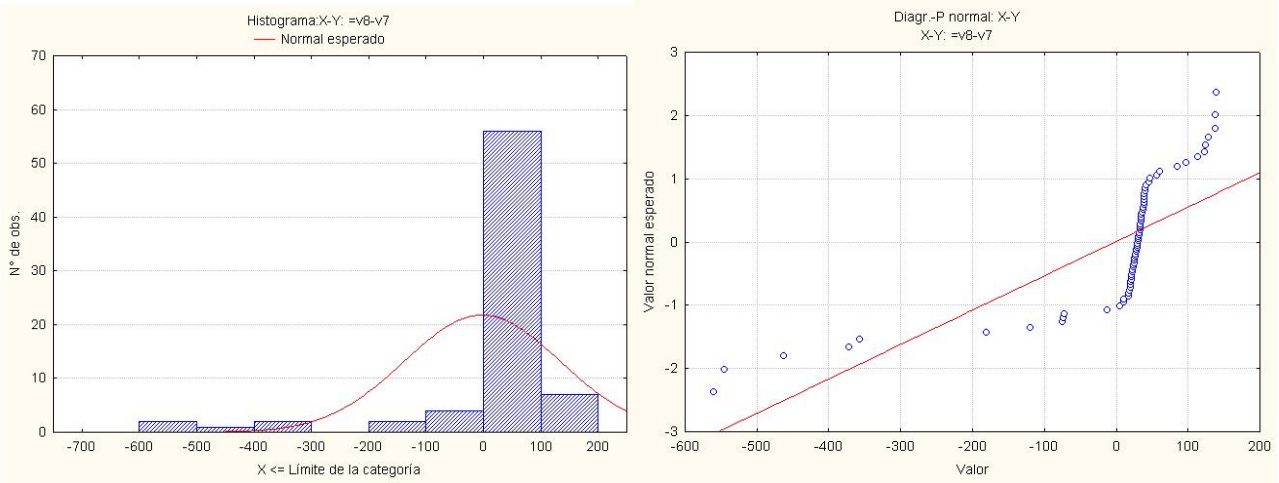
$$T_0 = \frac{\overline{X}_n - \overline{Y}_m}{S_{X-Y}/\sqrt{n}}$$

Y la conclusión es que, para un nivel de significación α=0,05, aceptamos H_0. De hecho, el *p*-valor es tan grande que "nos quedamos muy tranquilos" con la decisión que hemos tomado.

Nota: Este contraste que acabamos de estudiar, para muestras dependientes o pareadas, no es sino un caso particular del contraste para una muestra simple (el primer ejemplo que hemos visto), para la variable X-Y, cuya media poblacional es $\mu_{X-Y} = \mu_X - \mu_Y$, con valor de prueba 0.

$$\begin{cases} H_0 \equiv \mu_X = \mu_Y \\ H_1 \equiv \mu_X \neq \mu_Y \end{cases} \Leftrightarrow \begin{cases} H_0 \equiv \mu_{X\text{-}Y} = 0 \\ H_1 \equiv \mu_{X\text{-}Y} \neq 0 \end{cases}$$

EJERCICIO 2.6.- Realiza el contraste indicado para la variable DIFERENCIA (que es X-Y en el ejemplo que hemos estudiado) y comprueba que, en efecto, estamos haciendo lo mismo. "De paso", pégale un vistazo al histograma y al Diagrama de normalidad de DIFERENCIA.



Y, desde luego, si una impresión proporcionan los dos gráficos es que la muestra no proviene ni por asomo, de una población normal.

EJERCICIO 2.7.- Vamos a ver cómo andamos de lógica y si hemos entendido bien la dinámica de los contrastes de hipótesis paramétricos. Para la variable DIFERENCIA = X-Y = diferencia del volumen de negocio entre los años 2008 y 2007, nuestros estudios nos han llevado a concluir que:

I) Suponiendo que la población sigue una distribución normal en X-Y, ante las hipótesis

$$\begin{cases} H_0 \equiv \mu_{X\text{-}Y} = 0 \\ H_1 \equiv \mu_{X\text{-}Y} \neq 0 \end{cases}$$

hemos concluido que H_0 es CIERTA además, con un *p*-valor cercano a 1 (el acusado es INOCENTE; no hay prácticamente pruebas en su contra).

II) Los gráficos indican claramente que la población NO sigue una distribución normal en la variable X-Y.

Así, considerando I) y II), ¿ACEPTAMOS o RECHAZAMOS H_0? (¿declaramos al acusado INOCENTE o CULPABLE?).

Nota importante: Naturalmente también es posible plantearse el estudio de los contrastes de hipótesis unilaterales que quedan recogidos en el siguiente cuadro, para los que sirven los mismos análisis y comentarios que hicimos en el caso simple (en particular, los referidos al cálculo del p-valor a partir del calculado por el programa para el contraste bilateral).

TABLA 2.4: CONTRASTES DE HIPÓTESIS PARA DOS MUESTRAS

Situación	Hipótesis	Estadístico	R.A.
X, Y independientes X: $N(\mu_x,\ \sigma_x)$ Y: $N(\mu_y,\ \sigma_y)$ $\mu_x,\ \mu_y$ desconocidas $\sigma_x = \sigma_y$ desconocida $X_1, X_2, \ldots, X_n$ m.a. $Y_1, Y_2, \ldots, Y_m$ m.a.	$\begin{cases} H_0 \equiv \mu_x = \mu_y \\ H_1 \equiv \mu_x \neq \mu_y \end{cases}$ $\begin{cases} H_0 \equiv \mu_x \geq \mu_y \\ H_1 \equiv \mu_x < \mu_y \end{cases}$ $\begin{cases} H_0 \equiv \mu_x \leq \mu_y \\ H_1 \equiv \mu_x > \mu_y \end{cases}$	$T_0 = \dfrac{\overline{X}_n - \overline{Y}_m}{S_p\sqrt{\dfrac{1}{n}+\dfrac{1}{m}}}$ donde $S_p = \sqrt{\dfrac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}}$	$\left(-t_{n+m-2;\,\frac{\alpha}{2}},\ t_{n+m-2;\,\frac{\alpha}{2}}\right)$ $\left(-t_{n+m-2;\,\alpha},\ +\infty\right)$ $\left(-\infty,\ t_{n+m-2;\,\alpha}\right)$
X, Y relacionadas X-Y: $N(\mu_{x-y},\ \sigma)$ $\mu_x,\ \mu_y$ desconocidas $\mu_{x-y} = \mu_x - \mu_y$ σ desconocida $X_1, X_2, \ldots, X_n$ m.a. $Y_1, Y_2, \ldots, Y_n$ m.a. y, por lo tanto, $X_1 - Y_1, \ldots, X_n - Y_n$ m.a.	$\begin{cases} H_0 \equiv \mu_x = \mu_y \\ H_1 \equiv \mu_x \neq \mu_y \end{cases}$ $\begin{cases} H_0 \equiv \mu_x \geq \mu_y \\ H_1 \equiv \mu_x < \mu_y \end{cases}$ $\begin{cases} H_0 \equiv \mu_x \leq \mu_y \\ H_1 \equiv \mu_x > \mu_y \end{cases}$ equivalentemente $\begin{cases} H_0 \equiv \mu_{x-y} = 0 \\ H_1 \equiv \mu_{x-y} \neq 0 \end{cases}$ $\begin{cases} H_0 \equiv \mu_{x-y} \geq 0 \\ H_1 \equiv \mu_{x-y} < 0 \end{cases}$ $\begin{cases} H_0 \equiv \mu_{x-y} \leq 0 \\ H_1 \equiv \mu_{x-y} > 0 \end{cases}$	$T_0 = \dfrac{\overline{(X-Y)}_n}{S_{x-y}/\sqrt{n}}$	$\left(-t_{n-1;\,\frac{\alpha}{2}},\ t_{n-1;\,\frac{\alpha}{2}}\right)$ $\left(-t_{n-1;\,\alpha},\ +\infty\right)$ $\left(-\infty,\ t_{n-1;\,\alpha}\right)$

EJERCICIO 2.8.- Queremos estudiar el Volumen de Negocio de las empresas al por menor en la Comunidad Autónoma durante el año 2008 (variable X=NEGOCIO_08), nos piden:

a) Calcular un intervalo de confianza con la mayor precisión posible (entre los niveles de confianza habituales).

b) Sin realizar ningún cálculo, a la vista del intervalo obtenido en el apartado anterior, ¿qué decisión tomarías si te pidiesen contrastar las hipótesis:

$$\begin{cases} H_0 \equiv \mu = 300 \\ H_1 \equiv \mu \neq 300 \end{cases}$$

para un nivel de significación α=0'1?

c) Un representante del mundo empresarial afirma que la media de X=NEGOCIO_08, superó los 325 miles de euros. ¿Se sostiene dicha afirmación (para un nivel de significación α=0'05) con nuestros datos?

d) El mismo representante también indica que, considerando las empresas de Equipamiento personal y las de Equipamiento del hogar, éstas obtuvieron similares medias de Volumen de Negocio durante el año 2008. ¿Se sostiene dicha afirmación (para un nivel de significación α=0'05) con nuestros datos?

e) ¿Qué condiciones habrían de cumplirse, en cada uno de los apartados anteriores, para que los resultados obtenidos puedan ser considerados válidos?

EJERCICIO 2.9.- Un equipo médico realiza un estudio para comparar la eficacia de dos tratamientos en la mejora del nivel de colesterol. Para ello, selecciona una muestra de 8 parejas de gemelos: a un hermano de cada pareja aplica el Tratamiento 1 y al otro el 2. Los resultados son:

Mejora (%)	Pareja	Tratamiento
16	López	Tratamiento 1
25	García	Tratamiento 1
22	Martínez	Tratamiento 1
21	Gómez	Tratamiento 1
6	Fernández	Tratamiento 1
7	Rodríguez	Tratamiento 1
22	Hernández	Tratamiento 1
41	Sáenz	Tratamiento 1
10	López	Tratamiento 2
4	García	Tratamiento 2
8	Martínez	Tratamiento 2
7	Gómez	Tratamiento 2
17	Fernández	Tratamiento 2
5	Rodríguez	Tratamiento 2
10	Hernández	Tratamiento 2
21	Sáenz	Tratamiento 2

¿Puede deducirse, con un nivel de significación α=0'05, que ambos tratamientos son igual de efectivos? ¿Y si consideramos α=0'01? ¿Qué condiciones habrían de cumplirse para que los resultados pudiesen ser considerados válidos?

Indicación: Antes de nada, identificad de qué tipo son las muestras (¿relacionadas o independientes?) y pensad en cómo tenéis que escribirle los datos a STATISTICA para que lo entienda (no sirve reproducir la tabla tal y como os la hemos dado. ¡Tenemos que adaptarnos a los formatos y "manías" de los programas informáticos!).

ANEXO: INFERENCIA SOBRE UNA PROPORCIÓN POBLACIONAL DESCONOCIDA

Una situación muy habitual en la práctica consiste en estudiar una muestra de una población de la cual desconocemos la proporción, *p,* de una determinada característica (refirámonos a ella como éxito) de sus individuos. No nos costaría mucho esfuerzo desarrollar un estudio muy similar a los que hemos realizado en esta sesión. Por ejemplo:

1) Parece razonable estimar *p* por la proporción muestral, que suele denotarse $\hat{p}$, y naturalmente, se calcula como:

$$\hat{p} = \frac{\text{número de éxitos en la muestra}}{n}$$

2) A partir de una muestra aleatoria de tamaño ***n*** suficientemente grande, puede verse que un intervalo de confianza para *p*, con nivel de confianza 100x(1-α)%, es

$$IC_p((1-\alpha)\%) = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Nota: la exigencia: ***n*** "suficientemente grande", se debe a que en el desarrollo teórico se aplica un resultado de aproximación que lo requiere. Un criterio habitual para decidirlo es:

- si $\hat{p} \le 0'5$ ha de cumplirse que $n.\hat{p} \ge 5$
- si $\hat{p} > 0'5$ ha de cumplirse que $n.(1-\hat{p}) \ge 5$

Ejercicio: Determinar un intervalo de confianza, al 95%, para la proporción *p* de comercios que en 2007 tuvieron un volumen de negocio mayor a 300 miles de euros.

Solución: Desgraciadamente estamos ante un ejemplo que STATISTICA no resuelve automáticamente y nos vemos obligados a hacer las cuentas. Primero comprobamos que $\hat{p} = 40/76 = 0'5263$ y que $n.(1-\hat{p}) \ge 5$, y obtenemos,

$$IC_p(95\%) = 0'5263 \pm 1'96\sqrt{\frac{0'5263(1-0'5263)}{76}} = (0'414, 0'6385)$$

EJERCICIO 2.10.- Comprobad que no nos hemos equivocado al calcular $\hat{p}$, e interpretad el anterior resultado.

3) Podemos plantearnos y contrastar hipótesis, como, por ejemplo (suponiendo las condiciones del ejercicio anterior).

$$\begin{cases} H_0 \equiv p = 0.4 \\ H_1 \equiv p \neq 0.4 \end{cases}$$

Parece razonable tomar la decisión comparando el valor de prueba, p_0=0'4, y la proporción muestral $\hat{p} = 0'5263$.

Dado un nivel de significación α, un pequeño desarrollo teórico, suponiendo que ***n*** es suficientemente grande, lo que es habitual decidir con el criterio:

- si $p_0 \leq 0'5$ ha de cumplirse que $n.p_0 \geq 5$
- si $p_0 > 0'5$ ha de cumplirse que $n.(1-p_0) \geq 5$

la regla de decisión sería:

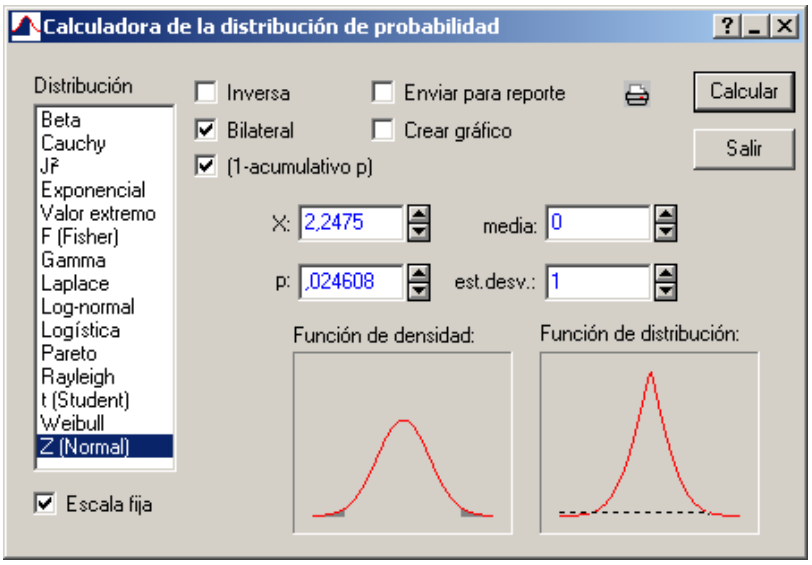
$$\text{ACEPTAR } H_0 \text{ sí y sólo sí} \qquad -Z_{\alpha/2} \leq \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} \leq Z_{\alpha/2}$$

Desgraciadamente, STATISTICA tampoco nos hace las cuentas:

$$T_0 = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0'5263-0'4}{\sqrt{0'4*(1-0'4)/76}} = 2'2475$$

Y utilizando la calculadora de STATISTICA, (en **Estadísticas/Calculadora probabilística/Distribuciones...**), podemos calcular las regiones críticas para el nivel de significación α deseado o, mucho mejor, el *p*-valor del contraste.

$$\frac{p}{2} = P(Z > T_0) = P(Z > 2'2475) = 0'0123 \Rightarrow p = 0'0246$$



¿Cuál será nuestra decisión para α=0'05? ¿Y para α=0'01?

EJERCICIO 2.11.- Calcular el p-valor para los contrastes unilaterales.

$$\begin{cases} H_0 \equiv p \geq 0.4 \\ H_1 \equiv p < 0.4 \end{cases} \qquad \begin{cases} H_0 \equiv p \leq 0.4 \\ H_1 \equiv p > 0.4 \end{cases}$$

¿Cuál será nuestra decisión para α=0'05?

TABLA 2.5: CONTRASTES PARA UNA PROPORCIÓN

Situación	Hipótesis	Estadístico	R.A.
p proporción poblacional desconocida n suficientemente grande	$\begin{cases} H_0 \equiv p = p_0 \\ H_1 \equiv p \neq p_0 \end{cases}$ $\begin{cases} H_0 \equiv p \geq p_0 \\ H_1 \equiv p < p_0 \end{cases}$ $\begin{cases} H_0 \equiv p \leq p_0 \\ H_1 \equiv p > p_0 \end{cases}$	$T_0 = \dfrac{\widehat{p} - p_0}{\sqrt{\dfrac{p_0\ (1-p_0)}{n}}}$	$\left(-Z_{\frac{\alpha}{2}},\ Z_{\frac{\alpha}{2}}\right)$ $\left(-Z_{\alpha},\ +\infty\right)$ $\left(-\infty,\ Z_{\alpha}\right)$

EJERCICIO 2.12.- Un representante del mundo empresarial señala que la proporción de comercios en los cuales se destruyó empleo (su plantilla disminuyó de 2008 a 2007) no superó el 35%. ¿Se sostiene dicha afirmación (para un nivel de significación α=0'05) con nuestros datos?

.

Comentario final:

En esta sesión hemos estudiado métodos de Inferencia Paramétrica que, para poder ser utilizados, requieren del cumplimiento previo de ciertas condiciones de aplicación. En la próxima sesión completaremos este estudio en dos direcciones:

- presentaremos herramientas para comprobar dichas condiciones. Notar que, en buena lógica, en el trabajo práctico, sería una tarea previa a la que hemos desarrollado.
- estudiaremos algunos métodos no paramétricos, que pueden emplearse en el estudio de situaciones similares a las enfrentadas en esta sesión. Son una alternativa a los paramétricos (obligada cuando éstos no pueden utilizarse).

Sesión 3. Inferencia no paramétrica

Al estimar los parámetros de un modelo se supone que los datos constituyen una muestra aleatoria de una distribución que, salvo por sus parámetros, es conocida.
La primera etapa de un estudio consiste, por tanto, en contrastar si las hipótesis básicas no están en contradicción con la muestra. Así, las primeras pruebas no paramétricas analizan:

- si las varianzas son iguales (homoscedasticidad)
- si la distribución supuesta es consistente con los datos, denominadas pruebas de bondad de ajuste: test de la χ^2, de Kolmogorov-Smirnov, de Shapiro-Wilk, ...

Por otra parte, ¿cuál es el efecto de un modelo distinto al supuesto?
El utilizar un modelo falso nos puede llevar a conclusiones no válidas o, aunque sean válidas, dejan de ser óptimas. Por ejemplo, los procedimientos que suponen normalidad llevan a estimaciones poco precisas cuando esta hipótesis no es válida, lo que se traduce en intervalos innecesariamente grandes o contrastes poco potentes. Por tanto, deberemos utilizar procedimientos alternativos que no requieran de suposiciones sobre la población salvo, en algunos casos, la de que sea continua.
Cuando tenemos una muestra utilizamos:

- contrastes de localización: test de los signos y de los rangos con signo de Wilcoxon

En el caso de querer contrastar diferencias entre dos muestras tenemos:

- dos muestras relacionadas: test de los signos y de los rangos con signo de Wilcoxon
- dos muestras independientes: test de rachas, de Kolmogorov-Smirnov y de la U de Mann-Whitney

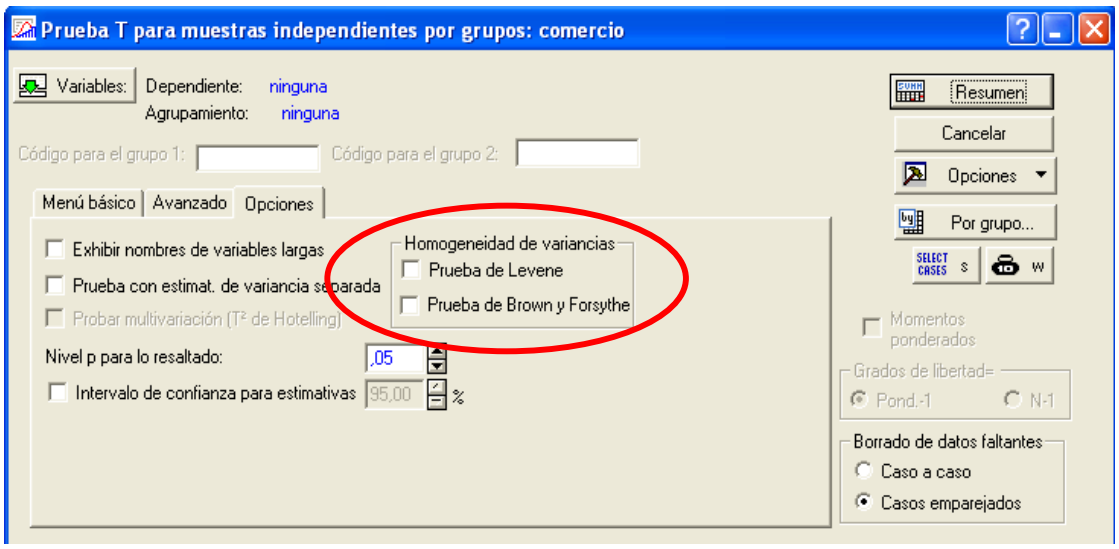
3.1. Pruebas de homoscedasticidad

En los test de igualdad de medias se supone homoscedasticidad es decir, que las varianzas poblacionales son iguales, afirmación que habrá que contrastar.

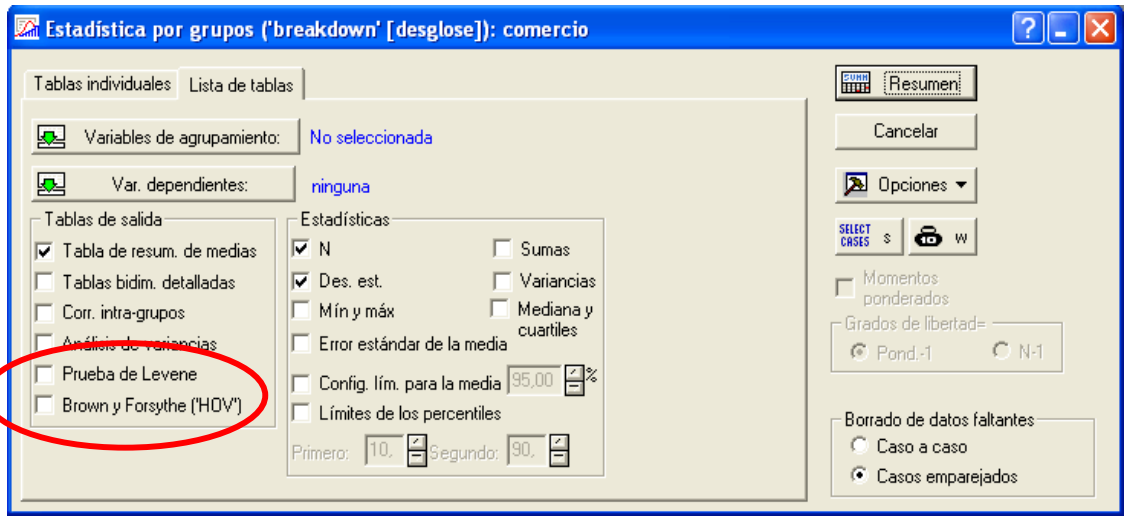
STATISTICA, además del contraste de razón de varianzas para aceptar la homoscedasticidad (*Ho*), proporciona dos contrastes más: el test de Levene y el test de Brown y Forsythe.
Estos últimos, a diferencia del de razón de varianzas, son robustos frente a desviaciones de la normalidad, es decir, continúan siendo válidos aunque no se verifique la normalidad.
Ambos test consisten en llevar a cabo un ANOVA utilizando como variable dependiente la diferencia en valor absoluto entre cada valor y la media (test de Levene) o la mediana (test de Brown y Forsythe) de su grupo, por lo que la distribución del estadístico será una F de Snedecor con grados de libertad del numerador igual al número de grupos menos 1 y grados de libertad del denominador tamaño de la muestra menos número de grupos.

STATISTICA aplica estos procedimientos desde varias opciones del menú:

- Cuando se desea realizar un contraste de dos medias para muestras independientes (**Estadísticas básicas y tablas / Prueba t independiente, por grupos**) se obtiene por defecto el contraste de la razón de varianzas a la vez del contraste de medias. Para obtener los otros test debemos ir a la pestaña Opciones y seleccionarlos. Con este procedimiento podemos contrastar la igualdad de sólo 2 varianzas.



- Cuando se quiere hacer un Análisis de la Varianza (**Estadísticas básicas y tablas / Análisis de variancia 'Breakdown' & de una vía**), pulsando la tecla de Lista de tablas, tenemos la opción de seleccionar la Prueba de Levene o la de Brown y Forsythe. Esta opción permite contrastar la igualdad de 2 o más varianzas.



EJERCICIO 3.1

¿Podemos aceptar que la dispersión (varianza) de la variable **negocio_07** es la misma entre los comercios que se ubican dentro de un centro comercial y los de fuera?

Prueba T; Agrupamien ubic: Ubicación (comercio)
Grupo 1: Fuera de cualquier centro comercial
Grupo 2: Dentro de un centro comercial

Variable	Media Fuera de cualquier centro comercial	Media Dentro de un centro comercial	Valor t	gl	p	N vál. Fuera de cualquier centro comercial	N vál. Dentro de un centro comercial
negocio_07	320,3872	317,0658	0,104305	74	0,917210	57	19

Desv. est. Fuera de cualquier centro comercial	Desv. est. Dentro de un centro comercial	Razón-F Variancias	p Variancias	Levene F(1,gl)	gl Levene	p Levene	Brn-Fors F(1,gl)	gl Brn-Fors	p Brn-Fors
123,1613	110,5064	1,242147	0,627655	0,078107	74	0,780660	0,061906	74	0,804198

EJERCICIO 3.2

¿Podemos aceptar que la dispersión (varianza) de la variable **negocio_07** es la misma entre los comercios de los distintos grupos de actividad?

Prueba de homogenidad de variancias de Levene (comercio)
Efectos marcados son significantes con p < ,05000

Variable	Sum.de Efecto	gl Efecto	Med.cu Efecto	Sum.de Error	gl Error	Med.cu Error	F	p
negocio_07	16718,90	3	5572,967	413303,8	72	5740,330	0,970844	0,411332

Prueba de Brown-Forsythe de Homolog. de variancias (comercio)
Efectos marcados son significantes con p < ,05000

Variable	Sum.de Efecto	gl Efecto	Med.cu Efecto	Sum.de Error	gl Error	Med.cu Error	F	p
negocio_07	17012,03	3	5670,678	442870,8	72	6150,984	0,921914	0,434712

EJERCICIO 3.3

¿Podemos aceptar que la dispersión (varianza) de la variable **negocio_08** es la misma entre los comercios que se ubican dentro de un centro comercial y los de fuera?, ¿y entre los comercio de los distintos grupos de actividad?

3.2. Pruebas de normalidad

Estos procedimientos tratan de averiguar si nuestros datos pueden proceder de una variable con distribución normal (*Ho*).

Contraste de la χ^2 de Pearson de bondad de ajuste

Los contrastes de la χ^2 comparan las frecuencias observadas frente a las esperadas con la hipótesis que se contrasta.

Para la aplicación del contraste es necesario que la muestra sea grande (mínimo 25), las observaciones deben estar agrupadas en clases (al menos 5 clases), que cada clase tenga al menos 5 datos y que la frecuencia esperada sea también de al menos 5.

El estadístico de la χ^2 tiene la siguiente expresión:

$\sum_i \frac{(O_i - E_i)^2}{E_i}$ que sigue una distribución χ^2

Donde: O_i y E_i son, para cada clase, las frecuencias observadas y esperadas respectivamente.

Cuando deseamos realizar un contraste de bondad de ajuste, el estadístico de la χ^2 de Pearson sigue una distribución asintótica χ^2 con $k - r - 1$ grados de libertad siendo k el número de clases y r el número de parámetros estimados del modelo. Es decir:

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \sim \chi^2_{k-r-1}$$

La aplicación más frecuente de este contraste para la normalidad es a problemas de muestra grande y donde μ y σ^2 se estiman a partir de los datos mediante $\bar{x}$ y s^2 por lo que la distribución del estadístico será de una χ^2_{k-3}.

Si la variable sigue una ley normal, se espera que no haya mucha diferencia entre la frecuencia observada y la esperada por lo que se rechaza la normalidad para valores grandes del estadístico.

Contraste de Kolmogorov – Smirnov

Este contraste de bondad de ajuste compara la probabilidad acumulada de la distribución teórica de una variable continua frente a la empírica mediante el estadístico D de Kolmogorov – Smirnov.

El estadístico de Kolmogorov – Smirnov es $D = sup_{x \in R} |F(x) - F_n(x)|$ siendo $F(x)$ y $F_n(x)$ las frecuencias relativas acumuladas teóricas y observadas respectivamente.

Los valores críticos para este contraste están tabulados en el supuesto de que no se requiera de la estimación de parámetros, sin embargo, cuando el contraste necesita de la estimación de parámetros, esta tabulación clásica conduce a un contraste muy conservador, es decir, tiende a aceptar la hipótesis nula.
Para mejorar el contraste, Lilliefors tabuló el estadístico D de Kolmogorov – Smirnov cuando estimamos los parámetros media (μ) y varianza (σ^2) de la distribución normal con sus valores muestrales $\overline{x}$ y s^2. Se rechaza la normalidad para valores grandes del estadístico *D*.

Contraste de Shapiro y Wilk

Este contraste mide el ajuste de la muestra representada en papel probabilístico normal a una recta. Se rechaza la normalidad cuando el ajuste es malo, que corresponde a valores pequeños de estadístico W de Shapiro – Wilk.

Contrastes de Asimetría y Curtosis

Los coeficientes de asimetría A y curtosis K son 0 bajo la hipótesis de normalidad por lo que valores alejados de 0 en esos coeficientes nos indican la posible no normalidad de los datos.
Para tamaños de muestra grande (n al menos 50) el coeficiente de asimetría A sigue una distribución normal de media 0 y varianza $6/n$.
Para tamaños de muestra grande (n al menos 200) el coeficiente de custosis K sigue una distribución normal de media 0 y varianza $24/n$.
Valores pequeños de estos estadísticos nos permiten aceptar la normalidad de los datos.

Métodos gráficos para contrastar la normalidad

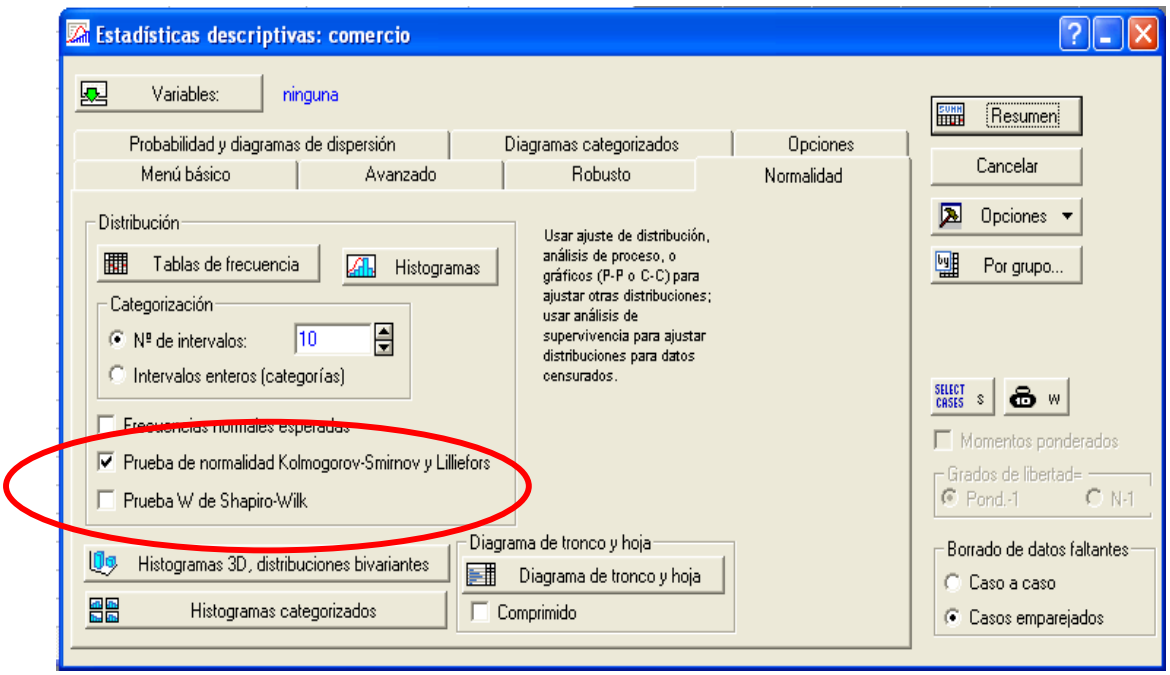
De la observación de la asimetría y curtosis de gráficos como el histograma, gráfico de tallos y hojas y diagrama de cajas podemos también darnos cuenta de la necesidad de no aceptar la normalidad en casos con una gran asimetría o una curtosis muy alejada de la de la curva normal.

Los gráficos de probabilidad normal para contrastar la normalidad. comparan la probabilidad acumulada observada (la que proporcionan los datos) frente a la probabilidad acumulada esperada (la obtenida de la distribución teórica a contrastar). Para poder aceptar la distribución teórica los puntos se agruparán en torno a una línea recta. El estadístico de Shapiro – Wilk es el coeficiente de determinación del ajuste de esos puntos a la recta por lo que valores grandes del estadístico (cerca de 1) nos llevan a aceptar la hipótesis de normalidad.

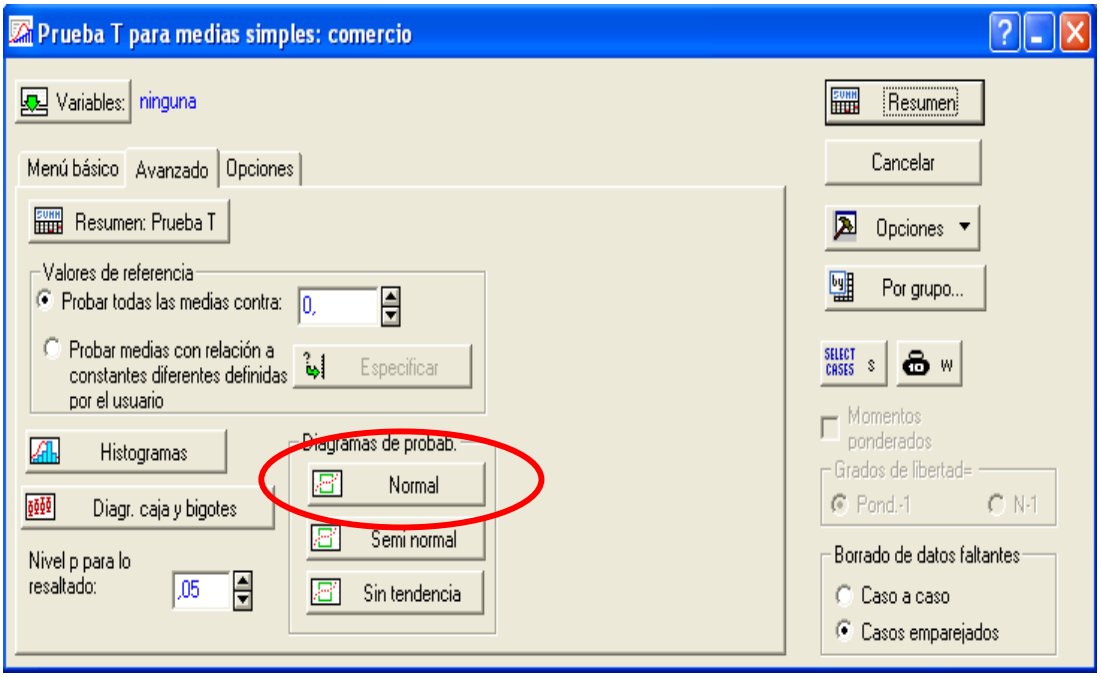
STATISTICA:

Podemos utilizar los distintos contrastes de normalidad desde varias opciones del menú:

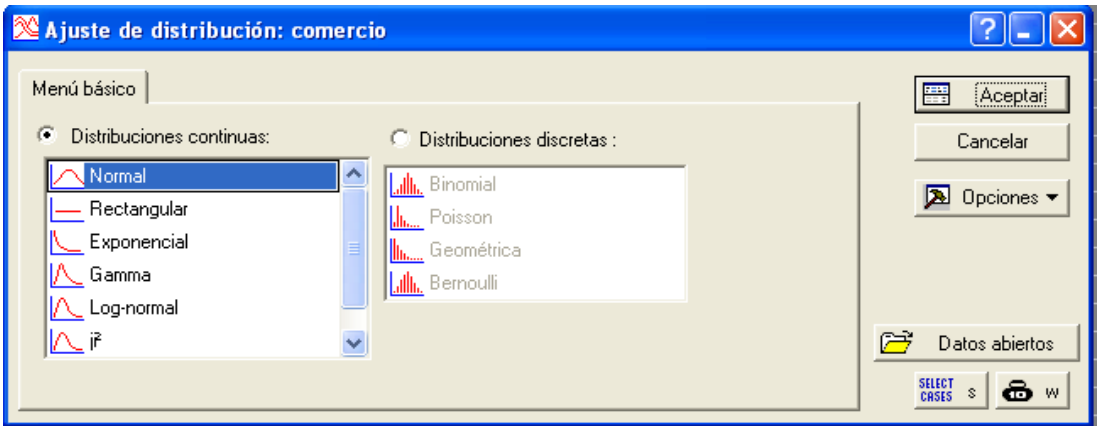
- **Estadísticas básicas y tablas / Estadísticas descriptivas** donde podemos obtener el contraste de Kolmogorov-Smirnov con y sin la corrección de Lilliefors y el contraste de Shapiro - Wilk

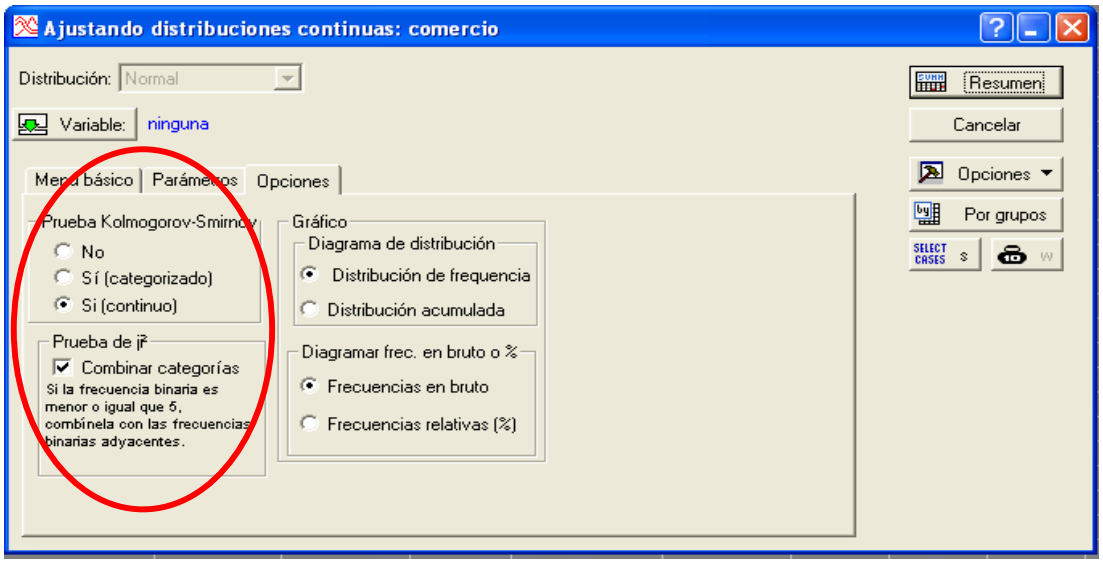


- **Estadísticas básicas y tablas / Prueba t, muestra simple** donde podemos obtener el diagrama de probabilidad de la normal aunque este contraste gráfico siempre será menos objetivo que uno basado en un estadístico con un p-valor.

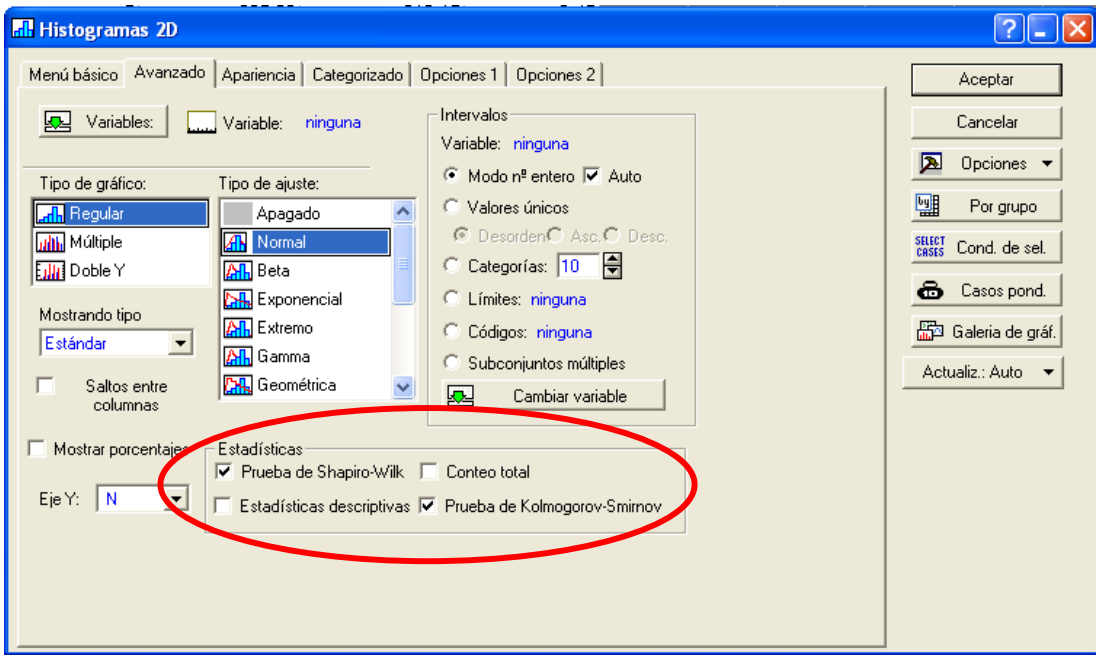


- **Ajuste de distribución** donde podemos obtener el contraste de Kolmogorov-Smirnov con y sin la corrección de Lilliefors y el contraste de la χ^2

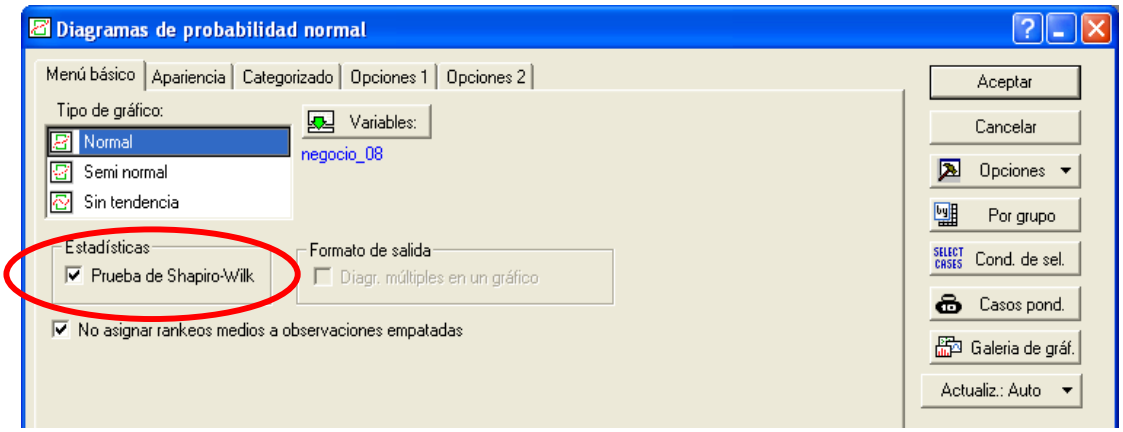




- **Gráficos / Histogramas** donde podemos obtener el contraste de Kolmogorov-Smirnov con y sin la corrección de Lilliefors y el de Shapiro - Wilk



- **Gráficos / Gráficos 2D / Diagramas de probabilidad normal** donde podemos obtener el contraste de Shapiro - Wilk



La exploración de los datos ayuda a determinar si son adecuadas las técnicas estadísticas que está teniendo en consideración para el análisis de los datos. Si no se puede aceptar la normalidad de estos, el usuario necesita utilizar pruebas no paramétricas.

EJERCICIO 3.4

Estudia la hipótesis de normalidad de la variable **negocio_08**.

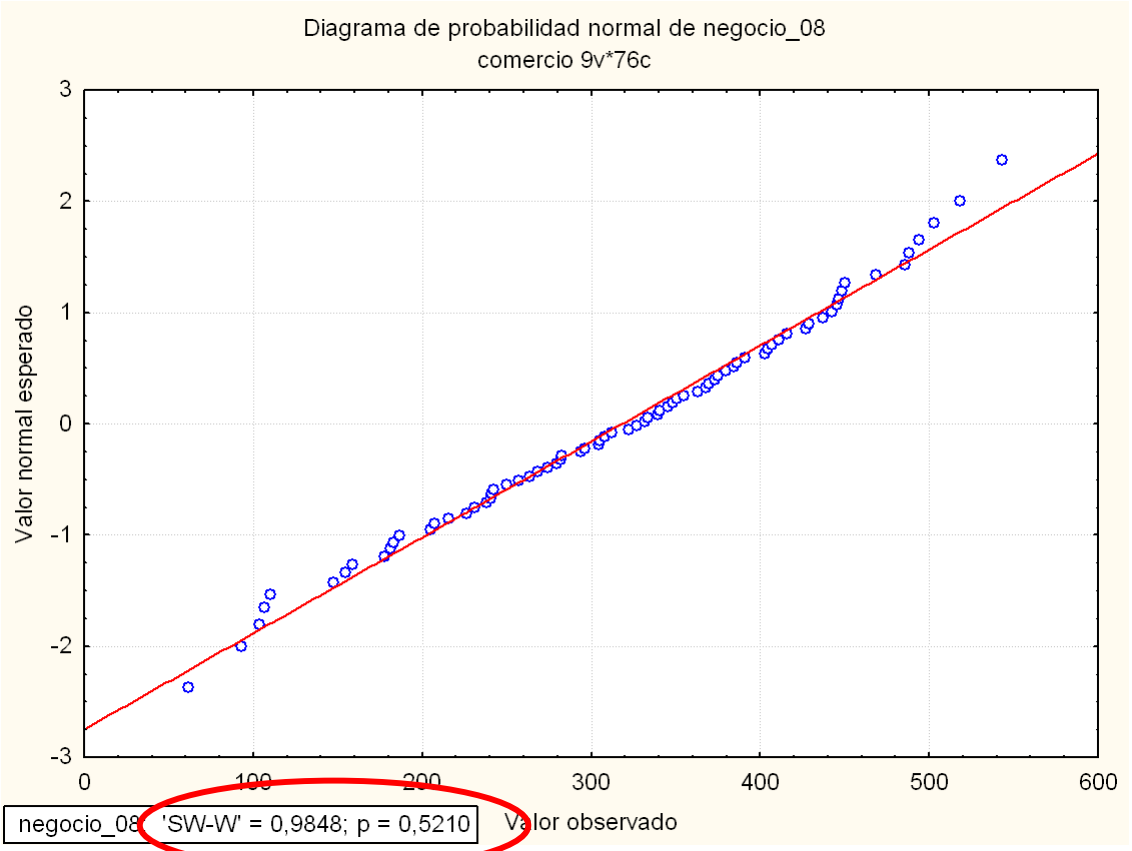
A la vista de los datos (media, media recortada, mediana y coeficiente de asimetría) podemos aceptar que la asimetría es relativamente pequeña.

Utilizando el contraste de asimetría tenemos un valor de $Z = \frac{-0'226}{\sqrt{6/76}} = -0'804$ lo que nos proporciona un $p-valor$ (unilateral) = 0'211, lo que nos permite aceptar la simetría de la distribución.

En cuanto a la curtosis, aunque el coeficiente nos indica que es menos apuntada que la curva normal, éste no es significativamente distinto de 0 ya que: $Z = \frac{-0'602}{\sqrt{24/76}} = -1'071$, por tanto, $p-valor$ (unilateral) = 0'142

También podemos observarlo con el histograma y el diagrama de cajas.

El diagrama de probabilidad normal, al que le añadimos el contraste de Shapiro – Wilk también nos permite aceptar la normalidad.



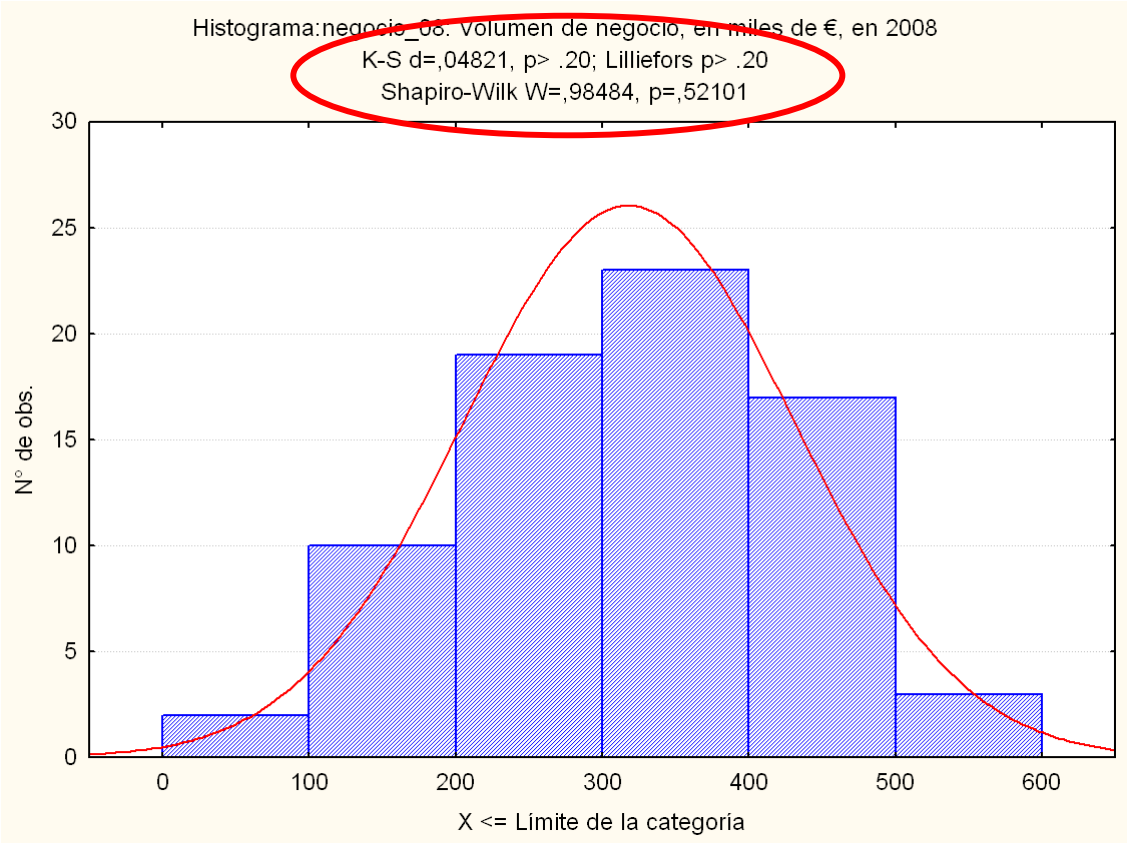
Si ajustamos la variable a la distribución normal y calculamos la tabla de las frecuencias observadas y esperadas obtenemos los contrastes de K-S y la χ^2.

Los dos contrastes nos permiten aceptar la hipótesis de normalidad de la distribución de la que proceden las observaciones de **negocio_08**. En ambos casos el *p – valor* es grande. (n.s. significa que es superior a 0,200, que es el máximo valor que está tabulado para el contraste de Kolmogorov – Smirnov con la corrección de Lilliefors).

Variable:negocio_08, Distribución: Normal (comercio)
Kolmogorov-Smirnov d =0,04821, p = n.s., Lilliefors p = n.s.
ji-cuadrado = 2,78722, gl=5(ajustada), p= 0,73275

Superior Límite	Observado Frecuencia	Acumul. de Observado	% Observado	Acumul. de % Observado	Esperad. Frecuencia	Acumul. de Esperad.	% Esperad.	Acumul. de % Esperad.	Observado-Esperad.
<= 50,00000	0	0	0,00000	0,0000	0,66372	0,66372	0,89693	0,8969	-0,66372
100,00000	2	2	2,70270	2,7027	1,34039	2,00412	1,81134	2,7083	0,65961
150,00000	4	6	5,40541	8,1081	3,09345	5,09756	4,18033	6,8886	0,90655
200,00000	6	12	8,10811	16,2162	5,89353	10,99109	7,96423	14,8528	0,10647
250,00000	10	22	13,51351	29,7297	9,26930	20,26039	12,52607	27,3789	0,73070
300,00000	9	31	12,16216	41,8919	12,03562	32,29601	16,26435	43,6433	-3,03562
350,00000	12	43	16,21622	58,1081	12,90171	45,19772	17,43475	61,0780	-0,90171
400,00000	11	54	14,86486	72,9730	11,41788	56,61560	15,42957	76,5076	-0,41788
450,00000	12	66	16,21622	89,1892	8,34218	64,95779	11,27322	87,7808	3,65782
500,00000	5	71	6,75676	95,9459	5,03178	69,98957	6,79970	94,5805	-0,03178
550,00000	3	74	4,05405	100,0000	2,50552	72,49508	3,38584	97,9663	0,49448
< Infinidad	0	74	0,00000	100,0000	1,50492	74,00000	2,03367	100,0000	-1,50492

Aunque ya hemos obtenido todos los contrastes de normalidad, vamos a completar el estudio con un histograma con los contrastes de Kolmogorov-Smirnov y Shapiro – Wilk.



Todos los contrastes nos llevan a aceptar la normalidad de la variable **negocio_08** por lo que aplicar un contraste paramétrico sobre su media será válido.

EJERCICIO 3.5
Estudia la hipótesis de normalidad de la variable que mide la diferencia de negocio (creada ya como **diferencia = negocio_08 – negocio_07**)

EJERCICIO 3.6
Estudia la hipótesis de normalidad de **pers_07** y **pers_08**. Etiqueta los casos mediante el código de identificación de la empresa (**ID**).

EJERCICIO 3.7
Estudia la hipótesis de normalidad de **negocio_07** por ubicación (**ubic**).

EJERCICIO 3.8
A la vista de los resultados obtenidos, ¿podemos aceptar que se verifican las condiciones para realizar un contraste sobre la media de la variable **pers_07**?

EJERCICIO 3.9
A la vista de los resultados obtenidos, ¿podemos aceptar que se verifican las condiciones para realizar un contraste sobre la diferencia de medias entre **negocio_07** y **negocio_08**?

3.3. Contrastes no paramétricos para una o dos muestras

Estos contrastes son la alternativa a los contrastes de una media o de diferencia de medias cuando la hipótesis de normalidad exigida para la aplicación del test de la t no se puede aceptar.

3.3.1. Contrastes para una muestra. Contrastes de localización

Estos contrastes tratan de analizar si la muestra puede procede de una población con una determinada medida de posición. La hipótesis nula conjetura que determinado percentil toma cierto valor. Nos vamos a centrar en el caso del percentil 50, es decir, la mediana. Tenemos, por tanto, el contraste:

$$\begin{cases} H_0 : Me = \theta \\ H_1 : Me \neq \theta \end{cases}$$

Prueba de los signos

Si aceptamos que la hipótesis nula es cierta y la mediana es θ, la probabilidad de que un elemento de la muestra sea superior a θ es ½, por tanto, en una muestra de tamaño n cabe esperarse que el número de valores superiores a θ sea de aproximadamente la mitad. Se rechazará la hipótesis nula si aparece un número excesivamente alto o excesivamente bajo respecto al 50%.

Los valores críticos de este contraste se determinan a partir de la distribución binomial puesto que el estadístico B que mide el número valores de la muestra superiores a θ sigue una ley binomial $B(n,½)$.

Además, en el caso de n grande, puede utilizarse la aproximación a la distribución normal por el teorema de Moivre.

Prueba de los rangos con signo de Wilcoxon

A diferencia del anterior, este contraste tiene en cuenta, no sólo el signo de las diferencias entre los valores de la muestra y la mediana que queremos contrastar, sino también, la magnitud de tales diferencias.

Este contraste utiliza los estadísticos T^+ y T^- de Wilcoxon que se obtienen de sumar los rangos, para las diferencias respecto a θ en valor absoluto, de los valores superiores o inferiores a θ respectivamente. Es decir, asignamos a cada elemento de la muestra, x_i, el rango, R_i, que obtenemos de ordenar los valores $|x_i - \theta|$ con lo que:

$$T^+ = \sum_{x_i > \theta} R_i \quad \text{y} \quad T^- = \sum_{x_i < \theta} R_i \, .$$

(para los empates se asigna la media de los rangos)

Si θ es la mediana cabe esperarse que T^+ y T^- sean aproximadamente iguales y, a su vez, aproximadamente iguales a la mitad de la suma de todos los rangos, con lo que, teniendo en cuenta que $T^+ + T^- = 1 + 2 + \ldots + n = \frac{1}{2}(1+n)n$, tanto T^+ como T^- deberían estar alrededor de $\frac{1}{4}(1+n)n$.

Para $n > 30$, los estadísticos de Wilcoxon siguen aproximadamente una ley normal de

$$\text{media} = \mu = \frac{n(n+1)}{4} \qquad \text{y} \qquad \text{varianza} = \sigma^2 = \frac{n(n+1)(2n+1)}{24}$$

3.3.2. Contrastes para 2 muestras relacionadas

Estos contrastes tratan de ver si dos muestras pueden proceder de la misma población o de poblaciones similares (*Ho*) cuando las muestras están relacionadas.

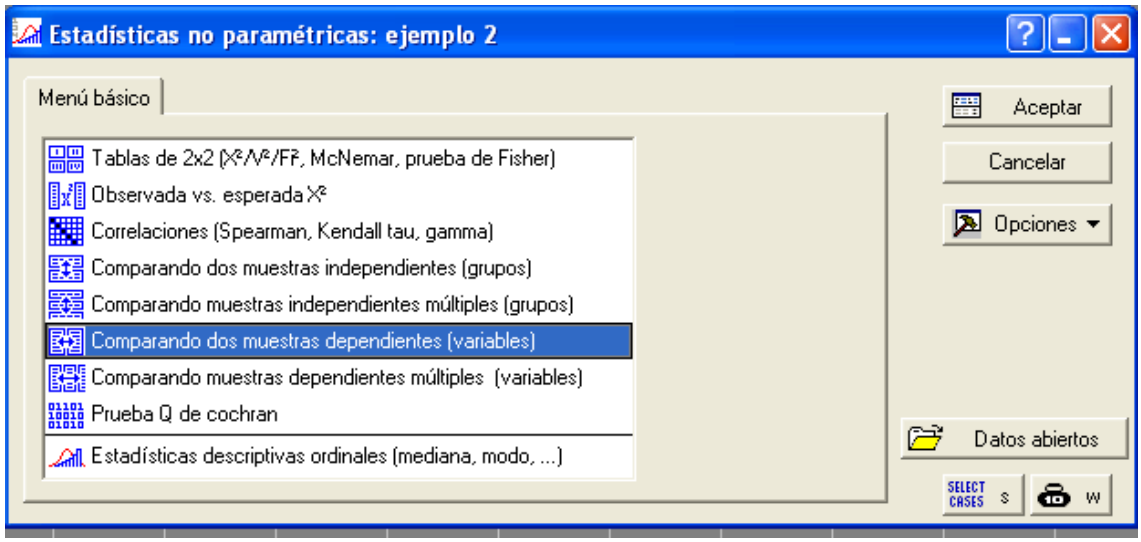
Dos de los contrastes para esta situación son: el test de los signos y el test de los rangos con signo de Wilcoxon, es decir, los que se utilizan en el caso de estar interesados en contrastar la mediana para una sola muestra.

Para poder utilizar estos contrastes deberemos obtener, para cada caso, las diferencias entre las dos muestras. Por tanto, contrastaremos como en los apartados 3.2.1.1 y 3.2.1.2 si la mediana de la diferencia es 0 ó no, es decir:

$$\begin{cases} H_0 : Me_{dif} = 0 \\ H_1 : Me_{dif} \neq 0 \end{cases}$$

STATISTICA: 2 muestras relacionadas

Tanto si decidimos realizar un contraste de localización de una muestra como si tenemos dos muestras relacionadas la opción es **Estadísticas / No paramétricos / Comparando 2 muestras dependientes (variables)**



Cuando queramos contrastar la mediana de una muestra deberemos crear una variable que contenga, en todos los casos, dicho valor a contrastar.

EJERCICIO 3.10

¿Podemos aceptar que la mediana de la variable que mide el volumen de negocio en 2007 es de 400.000€?, es decir, ¿la mitad de los comercios tienen un volumen de negocio inferior a 400.000€ y la otra mitad superior?

Primero debemos crear la variable que toma constantemente el valor 400, llamémosla **valor_400**. Después aplicamos los test de los signos y de Wilcoxon a las variables (relacionadas o dependientes) **negocio_07** y **valor_400**

	Prueba de signo (comercio) Pruebas marcadas son significantes para p <,05000			
Par de variables	N° de No-empat	% v < V	Z	nivel-p
negocio_07 & valor_400	76	78,94737	4,932438	0,000001

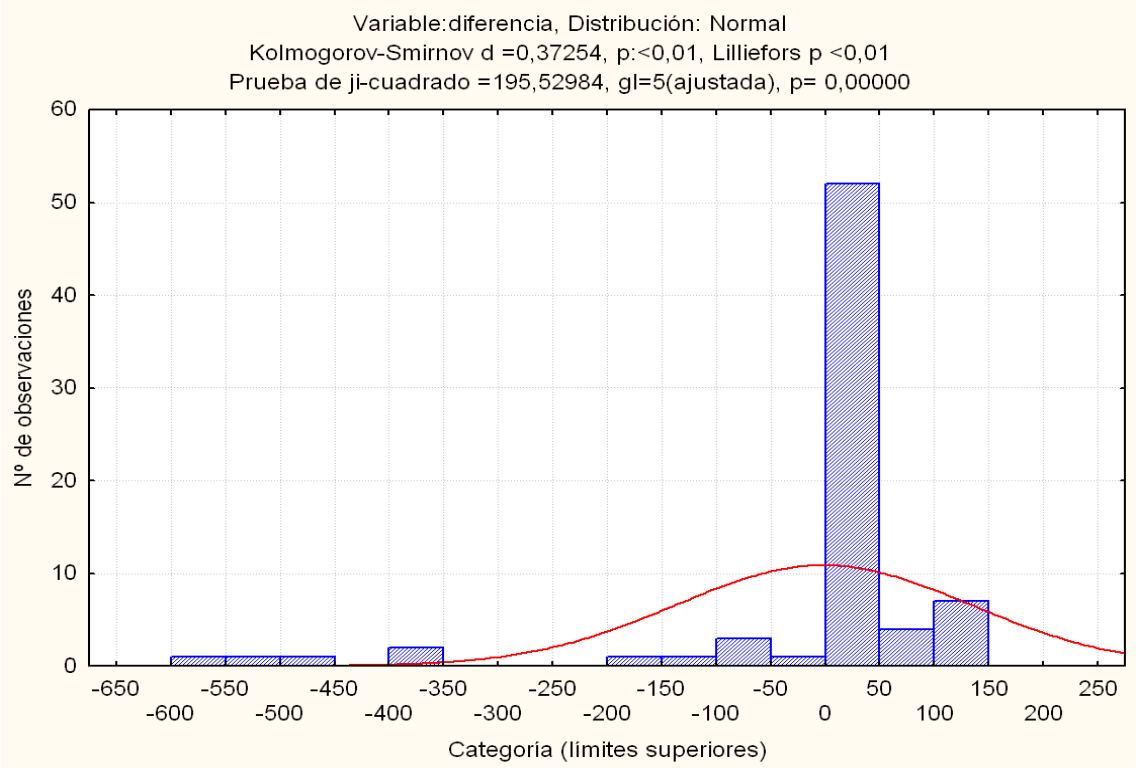
	Prueba de pares coincidentes de Wilcoxon (comercio) Pruebas marcadas son significantes para p <,05000			
Par de variables	Válido N	T	Z	nivel-p
negocio_07 & valor_400	76	475,0000	5,115224	0,000000

En ambos contrastes el *p*-valor nos indica que no podemos aceptar la hipótesis de que la mediana sea de 400.000€.

EJERCICIO 3.11

¿Podemos aceptar que el volumen de negocio de los comercios de la comunidad autónoma ha variado de 2007 a 2008?

Atendiendo a la pregunta nos piden que contrastemos si hay diferencias en cuanto al volumen de negocio de los comercios de la comunidad. Antes de decidirnos por un test paramétrico o no paramétrico debemos contrastar la normalidad de la variable diferencia de negocio 2007 a 2008 (creada ya y llamada **diferencia**).



El gráfico y los contrastes de Kolmogorov – Smirnov y la χ^2 nos indican que no se puede aceptar la normalidad de esta variable por lo que el contraste paramétrico de igualdad de medias no sería fiable. Optamos por lo tanto por la utilización de los no paramétricos: test de los signos y test de los rangos con signo de Wilcoxon.

	Prueba de signo (comercio) Pruebas marcadas son significantes para p <,05000				
Par de variables	N° de No-empat	% v < V	Z	nivel-p	
negocio_07 & negocio_08	74	85,13514	5,928630	0,000000	

	Prueba de pares coincidentes de Wilcoxon (comercio) Pruebas marcadas son significantes para p <,05000					
Par de variables	Válido N	T	Z	nivel-p		
negocio_07 & negocio_08	74	666,0000	3,886891	0,000102		

Ambos nos llevan a decir que los volúmenes de negocio en 2007 y 2008 presentan diferencias significativas (*p*-valores casi 0).

El contrastar las condiciones de validez cuando se va a aplicar un contraste paramétrico es de suma importancia. En este ejemplo concreto, si decidimos obviar que no se cumple la normalidad y aplicamos el contraste de la t para igualdad de medias obtenemos:

	Prueba t para muestras dependientes (comercio) Diferencias marcadas son significantes con p < ,05000							
Variable	Media	Des. est	N	Dif.	Des. est Dif.	t	gl	p
negocio_07	319,8534	120,8128						
negocio_08	318,1276	113,2829	74	1,725811	135,6491	0,109444	73	0,913151

lo que, ERRONEAMENTE, nos llevaría a aceptar que el volumen de negocio en 2007 es similar al de 2008.

EJERCICIO 3.12

¿Podemos aceptar que el volumen de negocio de los comercios ubicados en un centro comercial ha variado de 2007 a 2008?

EJERCICIO 3.13

¿Podemos aceptar que la cantidad de personal contratado por los comercios ha variado de 2007 a 2008?

EJERCICIO 3.14

¿Podemos aceptar que la cantidad de personal contratado por los comercios dedicados a alimentación ha variado de 2007 a 2008?

3.3.3. Contrastes para 2 muestras independientes

Estos contrastes tratan de estudiar si dos muestras pueden proceder de la misma población o de poblaciones similares (*Ho*) cuando las muestras son independientes.
Los contrastes que implementa STATISTICA son: el test de Mann-Whitney-Wilcoxon, el de Kolmogorov-Smirnov y el de rachas

Contraste de la U de Mann-Whitney-Wilcoxon

Este test se aplica a variables de tipo continuo y utiliza el estadístico de la U de Mann – Whitney que mide, de entre el total de pares formados entre un elemento de cada muestra, el número de ellos en los que el valor de la primera muestra es inferior al valor de la segunda.
Si aceptamos que la hipótesis nula de igualdad entre las muestras es cierta, la probabilidad de que un valor de la primera muestra sea inferior a otro de la segunda es ½. Si tenemos muestras de tamaño n_1 y n_2, el total de pares formados entre un elemento de cada muestra es $n_1 n_2$ y, por tanto, el número esperado de pares en los que el valor de la primera muestra es inferior al valor de la segunda es $n_1 n_2 / 2$. Rechazaremos H_0 para valores excesivamente grandes o pequeños del estadístico.

Si n_1 ó $n_2 > 20$, una generalización del Teorema del Límite Central nos permite aproximar el estadístico U a una distribución Normal $N(\mu, \sigma)$ con

$$\mu = \frac{n_1 n_2}{2} \quad \text{y} \quad \sigma^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Contraste de Kolmogorov-Smirnov para 2 muestras

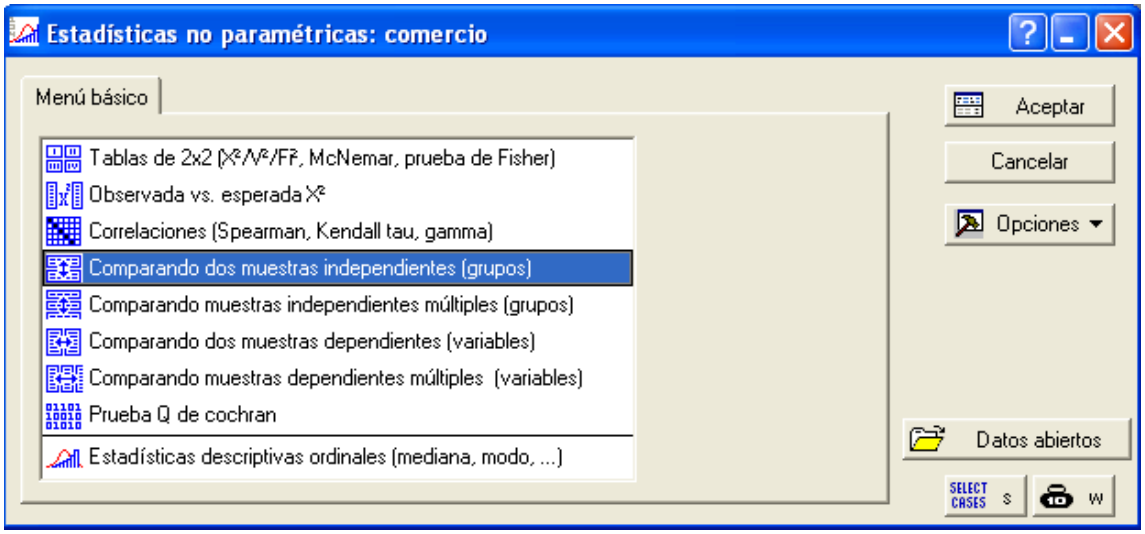
Este test se aplica a variables de tipo continuo y utiliza también el estadístico D de Kolmogorov-Smirnov comparando las frecuencias relativas acumuladas de las dos muestras.
El estadístico de Kolmogorov – Smirnov es $D = \sup_{x \in R} \left| F_{n_1}(x) - F_{n_2}(x) \right|$
siendo $F_{n_1}(x)$ y $F_{n_2}(x)$ las frecuencias relativas acumulada observadas para las dos muestras.
Si la hipótesis nula es cierta, la diferencia entre $F_{n_1}(x)$ y $F_{n_2}(x)$ no será muy grande por lo que se rechaza *Ho* para valores grandes del estadístico.

Contraste de Rachas de Wald-Wolfowitz

Este test contrasta si es aleatorio el orden de aparición de los valores de las dos muestras si éstos están ordenados. Un número excesivamente grande o excesivamente pequeño de rachas (conjunto de elementos consecutivos de la misma muestra) sugiere que las muestras no son similares, lo que nos lleva a rechazar la hipótesis nula.

STATISTICA: 2 muestras independientes

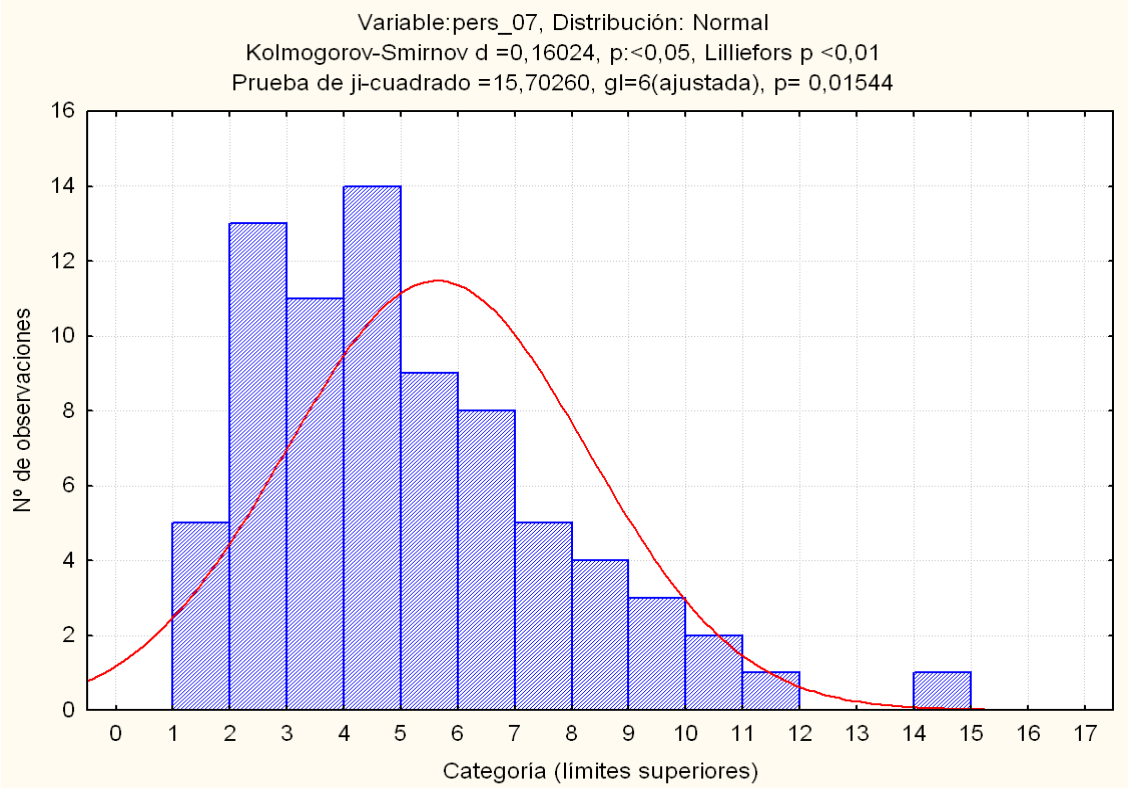
Para contrastar diferencias entre dos muestras independientes la opción es **Estadísticas / No paramétricos / Comparando 2 muestras independientes (grupos)**



EJERCICIO 3.15

¿Podemos aceptar que durante 2007 hubo un número de personas contratadas por comercio diferente según su ubicación?

Para responder debemos realizar un test para 2 muestras independientes, ¿paramétrico o no paramétricos? Contrastaremos primero la normalidad de la variable **pers_07**.



No podemos aceptar la normalidad por lo que nos decidimos a aplicar los contrastes no paramétricos:

Prueba de corridas de Wald-Wolfowitz (comercio)
Por variable ubic
Pruebas marcadas son significantes para p <,05000

Variable	N vál. Fuera de cualquier centro comercial	N vál. Dentro de un centro comercial	Media Fuera de cualquier centro comercial	Media Dentro de un centro comercial	Z	nivel-p	Ajustada por Z	nivel-p	N° de Corridas	N° de empates
pers_07	57	19	5,456140	6,157895	0,154672	0,877080	0,00	1,000000	30	26

Prueba Kolmogorov-Smirnov (comercio)
Por variable ubic
Pruebas marcadas son significantes para p <,05000

variable	Neg Máx Diferenc	Pos máx Diferenc	nivel-p	Media Fuera de cualquier centro comercial	Media Dentro de un centro comercial	Desv. es Fuera de cualquier centro comercial	Desv. es Dentro de un centro comercial	N vál. Fuera de cualquier centro comercial	N vál. Dentro de un centro comercial
pers_07	-0,175439	0,035088	p > .10	5,456140	6,157895	2,679753	2,522414	57	19

Prueba U de Mann-Whitney (comercio)
Por variable ubic
Pruebas marcadas son significantes para p <,05000

variable	Suma de Fuera de cualquier centro comercial	Suma de Dentro de un centro comercial	U	Z	nivel-p	Z ajustada	nivel-p	N vál. Fuera de cualquier centro comercial	N vál. Dentro de un centro comercial	2*1lado p exacto
pers_07	2090,000	836,0000	437,0000	-1,24757	0,212190	-1,25879	0,208108	57	19	0,214122

Vistos los *p*-valores, estos nos permiten aceptar que la ubicación del comercio no lleva a diferencias significativas en cuanto al número de personas contratadas por comercio en 2007.

EJERCICIO 3.16

¿Podemos aceptar diferencias significativas, durante 2007, en el volumen de negocio de los comercios dedicados a la alimentación y el resto?

EJERCICIO 3.17

¿Podemos aceptar que durante 2008 hubo un número de personas contratadas por comercio diferente según su ubicación?

EJERCICIO 3.18

¿Podemos aceptar que, durante 2008, en el volumen de negocio es similar entre los comercios dedicados a la alimentación y el resto?

3.4. Tratamiento de datos cualitativos. Test de la χ^2

Continuando con el estudio de la relación entre dos variables, vamos a analizar el caso en el que las variables que interesa relacionar son cualitativas. Esta situación aparece, por ejemplo, en el análisis de encuestas, en las que la mayoría de las variables que se recogen son opiniones o clasificaciones que no admiten fácilmente una expresión cuantitativa (partido al que se piensa votar, profesión, sexo, estado civil, actividades de ocio, etc.). Para estas variables no sirven los métodos de regresión, los cuales exigen que las variables sean cuantitativas, para ellas utilizaremos el test de la χ^2.

Contraste de la χ^2 de Pearson

La hipótesis nula que queremos contrastar es la de independencia entre dos variables cualitativas. Para esto, se construye lo que se denomina **Tabla de Contingencia** que es una tabla de doble entrada que recoge las frecuencias de las dos variables de forma conjunta, es decir, el número de veces que se ha observado cada combinación de valores de las dos variables.

Recordemos que el test de la χ^2 compara las frecuencias observadas frente a las esperadas con la hipótesis que se contrasta. También, que para su aplicación es necesario que la muestra sea grande (mínimo 25), que cada clase tenga al menos 5 observaciones y que la frecuencia esperada sea también de al menos 5.

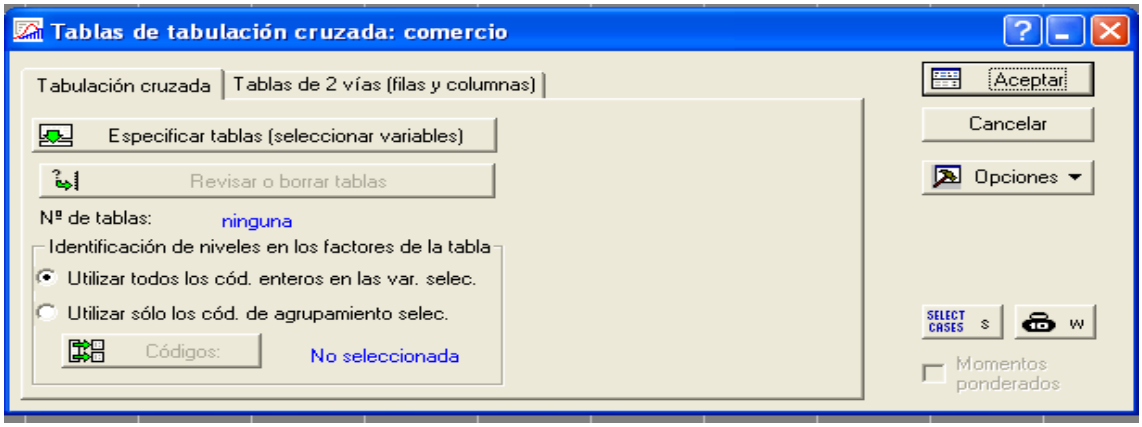
Para contrastes de independencia, el estadístico de la χ^2 de Pearson sigue una distribución asintótica χ^2 con $(c - 1)(d - 1)$ grados de libertad siendo c y d el número valores distintos de cada variable. Es decir:

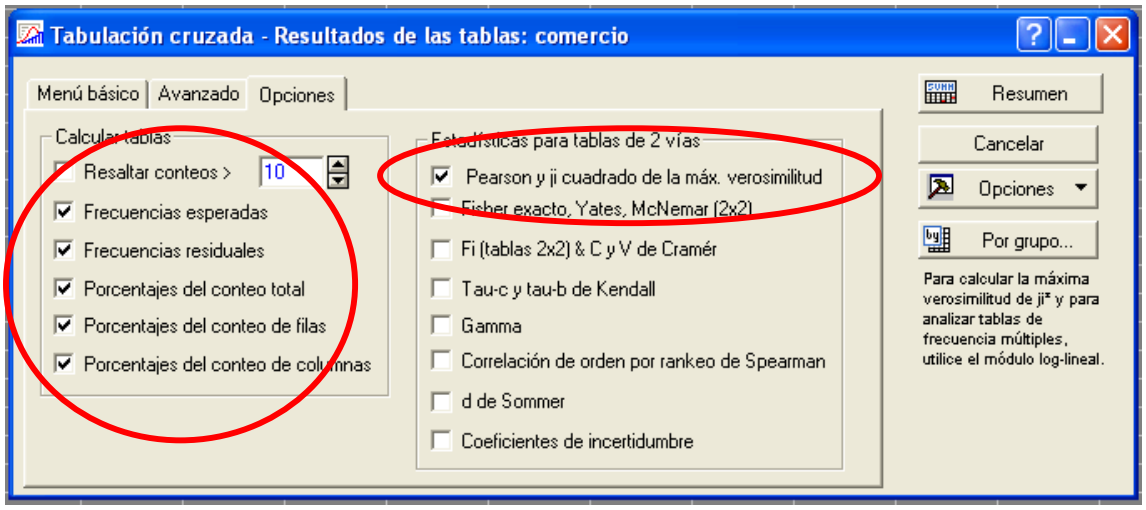
$$\sum_{i=1}^{c}\sum_{j=1}^{d}\frac{\left(n_{ij}-n_{i\bullet}n_{\bullet j}\right)^2}{n_{i\bullet}n_{\bullet j}} \sim \chi^2_{(c-1)(d-1)}$$

Si las variables son independientes, se espera que no haya mucha diferencia entre la frecuencia observada y la esperada por lo que se rechaza la independencia para valores grandes del estadístico.

STATISTICA:

Para analizar la posible relación de 2 variables cualitativas seleccionamos **Estadísticas / Estadísticas básicas y tablas / Tablas** donde podemos obtener varias tablas de frecuencias observadas y esperadas junto con contrastes de independencia.




Tabulación cruzada - Resultados de las tablas: comercio
Menú básico
Avanzado
Opciones
Calcular tablas
Resaltar conteos >
10
Frecuencias esperadas
Frecuencias residuales
Porcentajes del conteo total
Porcentajes del conteo de filas
Porcentajes del conteo de columnas
Estadísticas para tablas de 2 vías
Pearson y ji cuadrado de la máx. verosimilitud
Fisher exacto, Yates, McNemar (2x2)
Fi (tablas 2x2) & C y V de Cramér
Tau-c y tau-b de Kendall
Gamma
Correlación de orden por rankeo de Spearman
d de Sommer
Coeficientes de incertidumbre
Resumen
Cancelar
Opciones
Por grupo...
Para calcular la máxima verosimilitud de ji² y para analizar tablas de frecuencia múltiples, utilice el módulo log-lineal.

EJERCICIO 3.19

Analizar si el grupo de actividad de un comercio es independiente de que esté ubicado dentro o fuera de un centro comercial.

Para responder a esta pregunta solicitamos las tablas y el contraste de la χ^2.

Resumen tabla de frecuencia (comercio) n_{ij}
Tabla: grupo(4) x ubic(2)

grupo	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales $n_{i\bullet}$
Alimentación	6	31	37
Equipamiento personal	3	12	15
Equipamiento del hogar	6	4	10
Otros	4	10	14
Todos lo $n_{\bullet j}$	19	57	76

Tabla resumen :Frecuencias esperada (comercio) $n_{i\bullet}n_{\bullet j}$
Pearson ji-cuadrado:8,35109, gl=3, p=,039292

grupo	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales $n_{i\bullet}$
Alimentación	9,25	27,75	37,00
Equipamiento personal	3,75	11,25	15,00
Equipamiento del hogar	2,50	7,50	10,00
Otros	3,50	10,50	14,00
Todos lo $n_{\bullet j}$	19,00	57,00	76,00

Observando estas dos tablas vemos que existen diferencias entre las frecuencias observadas y las esperadas pero, ¿hasta qué punto estas diferencias son debidas al azar?

Tabla resumen :%s del total de columnas (comercio)
Tabla: grupo(4) x ubic(2)

grupo	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales
Alimentación	31,58%	54,39%	48,68%
Equipamiento personal	15,79%	21,05%	19,74%
Equipamiento del hogar	31,58%	7,02%	13,16%
Otros	21,05%	17,54%	18,42%
Todos lo	25,00%	75,00%	

Tabla resumen :%s del total de filas (comercio)
Tabla: grupo(4) x ubic(2)

grupo	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales
Alimentación	16,22%	83,78%	48,68%
Equipamiento personal	20,00%	80,00%	19,74%
Equipamiento del hogar	60,00%	40,00%	13,16%
Otros	28,57%	71,43%	18,42%
Todos lo	25,00%	75,00%	

Estas tablas también nos muestran diferencias, por ejemplo, el equipamiento para el hogar, en los centros comerciales es un 31,58% de los comercios mientas que fuera de

un centro comercial es un 7,02% de los comercios. Ahora bien, nos planteamos la misma pregunta ¿esta relación es significativa?

Tabla resumen :Frecuencias esperadas observadas mí (comercio)
Pearson ji-cuadrado:8,35109, gl=3, p=,039292

grupo	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales
Alimentación	-3,25000	3,25000	0,00
Equipamiento personal	-0,75000	0,75000	0,00
Equipamiento del hogar	3,50000	-3,50000	0,00
Otros	0,50000	-0,50000	0,00
Todos lo	0,00000	0,00000	0,00

Para el nivel de significación estándar de 5%, tenemos que el *p*-valor es inferior y por tanto debemos rechazar *H*o, concluyendo que las variables grupo de actividad y ubicación no son independientes, es decir, la relación que observábamos en las tablas ha resultado ser significativa para el nivel del 5%.

Por otra parte, tenemos un problema con las condiciones de aplicación de la prueba, recordemos que las frecuencias observadas y esperadas para cada clase deben ser de al menos 5. Para evitar este problema que cuestiona la validez del test, un procedimiento que podemos seguir es el de rehacer la tabla agrupando categorías. En nuestro ejemplo, la variable ubicación no se puede agrupar más porque ya está en el mínimo de dos categorías, pero para la variable grupo de actividad podríamos reducir sus cuatro categorías a dos: Alimentación-Equipamiento personal y Equipamiento del hogar-Otros.

La tabla de frecuencias esperadas, con la recodificación, es:

Tabla resumen :Frecuencias esperada (comercio)
Pearson ji-cuadrado:5,19658, gl=1, p=,022633

grupo2	ubic Dentro de un centro comercial	ubic Fuera de cualquier centro comercial	Fila Totales
Alimentación y eq. personal	13	39	52
Eq. del hogar y otros	6	18	24
Todos lo	19	57	76

Ahora, la frecuencia mínima esperada es 6, el *p*-valor es 0,023, lo que conduce a rechazar la independencia entre grupo de actividad recodificado y ubicación, y ya no tenemos motivos para cuestionar la validez de la prueba.

EJERCICIO 3.20

Analizar si el grupo de actividad de un comercio es independiente de su régimen jurídico.

Sesión 4. Análisis de la Varianza y Regresión Lineal

4.1. Análisis de la Varianza

En esta sección vamos a estudiar una técnica estadística denominada **Análisis de la Varianza (ANOVA)**. En la sesión 2 vimos la comparación de dos grupos respecto a una variable cuantitativa, o, equivalentemente, el estudio de la relación entre una variable cuantitativa y una variable (cualitativa o cuantitativa) que define dos grupos. Recordad el ejemplo: estudiar si el volumen de negocio es distinto según que los comercios estén ubicados dentro o fuera de centros comerciales, es equivalente a estudiar si la variable **ubicación,** que define dos grupos, está relacionada con la variable **volumen de negocio,** que es cuantitativa.

La cuestión que nos planteamos ahora es la de generalizar a un número cualquiera de grupos, es decir, comparar dos o más grupos respecto a una variable cuantitativa. Los grupos a comparar estarán definidos por los valores de alguna variable (cualitativa o cuantitativa), y por tanto, el planteamiento es equivalente al de estudiar la relación entre una variable que define varios grupos y una variable cuantitativa. Supongamos, por concretar, que estamos interesados en comparar los cuatro grupos de actividad (variable **grupo:** alimentación, equipamiento personal, equipamiento del hogar, otros) respecto al volumen de negocio en el año 2007 **(negocio_07).** Es decir, queremos saber si el grupo de actividad influye en el volumen de negocio, o, dicho de otra manera, si la variable **negocio_07,** que es cuantitativa, está relacionada con la variable **grupo,** que es cualitativa y define 4 grupos.

El problema de las comparaciones múltiples

Es posible que algún "experto" nos diga, a este respecto, que si ya hemos aprendido a comparar dos grupos, podemos comparar "los que nos echen": de dos en dos. Sería un camino más o menos largo según el número de grupos a comparar, pero teniendo acceso a un ordenador esto no supondría ninguna dificultad. Sin embargo, este procedimiento, aparte de no ser muy elegante, sería metodológicamente incorrecto.

Ello es debido a que en cada una de las comparaciones de dos grupos nos estamos exponiendo a un error (podemos concluir que los dos grupos son distintos, cuando en realidad no lo son) con una probabilidad dada por el nivel de significación que utilicemos. Si en cada una de *n* comparaciones independientes la probabilidad de equivocarse es 0'05, la probabilidad de equivocarse en alguna de ellas es $1 - (0'95)^n$, cantidad que, incluso para valores moderados de *n*, es bastante más grande que 0'05.

EJERCICIO 4.1.- Calcula el valor de esta probabilidad de equivocarse para $n = 2, 5, 10$ y 15.

En nuestro ejemplo, en el que tenemos 4 grupos a comparar, habría que hacer 6 comparaciones por parejas, lo que da una probabilidad de equivocarse en alguna de ellas igual a 0'2649. Para entendernos, si llevamos 6 cántaros a la fuente, la probabilidad de que alguno (uno o más) acabe roto es 0'2649, aunque la probabilidad de que se rompa

cada uno de ellos sea 0'05. En definitiva, y usando términos más técnicos, cuando se llevan a cabo comparaciones múltiples, el nivel de significación global puede ser mucho más elevado que el nivel de significación nominal de cada una de las comparaciones individuales. Como resultado, estaríamos asumiendo una probabilidad de equivocarnos demasiado elevada. Si el grupo de actividad no influye en el volumen de negocio, la probabilidad de que concluyamos erróneamente que sí influye sería 0'2649.

EJERCICIO 4.2.- Calcula el nivel de significación global si se comparan 6 grupos dos a dos con un nivel de significación en cada prueba igual a 0'10.

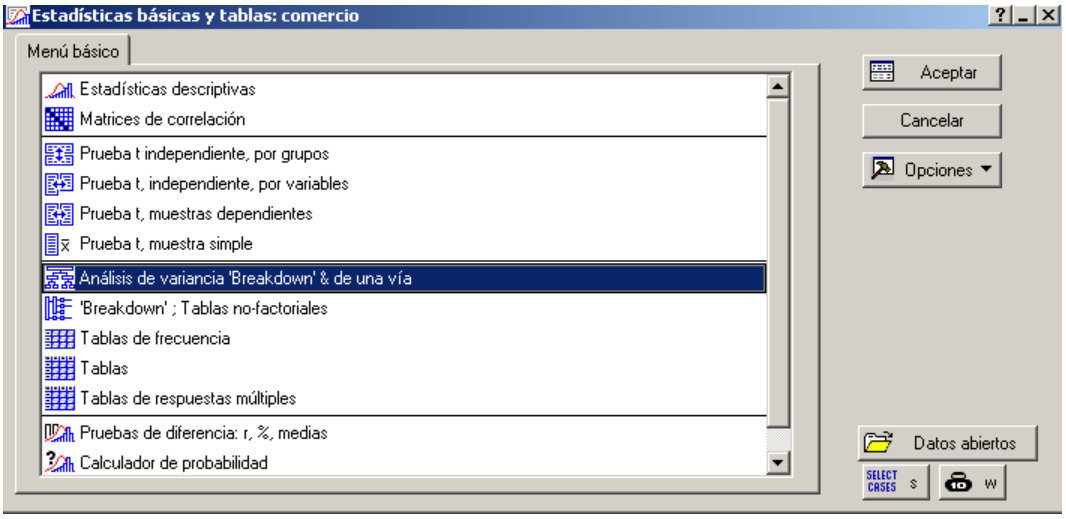
Muchos investigadores no parecen muy conscientes de las implicaciones de este problema, que puede manifestarse de distintas maneras. Así, si en un estudio hemos medido muchas variables y, a la hora del análisis estadístico, relacionamos, de dos en dos, todas con todas, sería raro que no encontráramos alguna relación significativa. Las relaciones que surgen en un contexto de comparaciones múltiples, especialmente las que no cabía esperar a priori, deben ser puestas en cuarentena, a la espera de que estudios posteriores las confirmen (o no). Esta situación se corresponde en cierta forma con lo que se llama, en probabilidad, el *"problema de las coincidencias"*: las coincidencias no deben sorprendernos, porque lo raro sería que no se produjese alguna.

Estas consideraciones nos sirven para justificar la necesidad de disponer de un procedimiento que nos permita eludir el problema de las comparaciones múltiples de *k* grupos dos a dos. Esta técnica se denomina Análisis de la Varianza (ANOVA).

El Análisis de la Varianza (ANOVA)

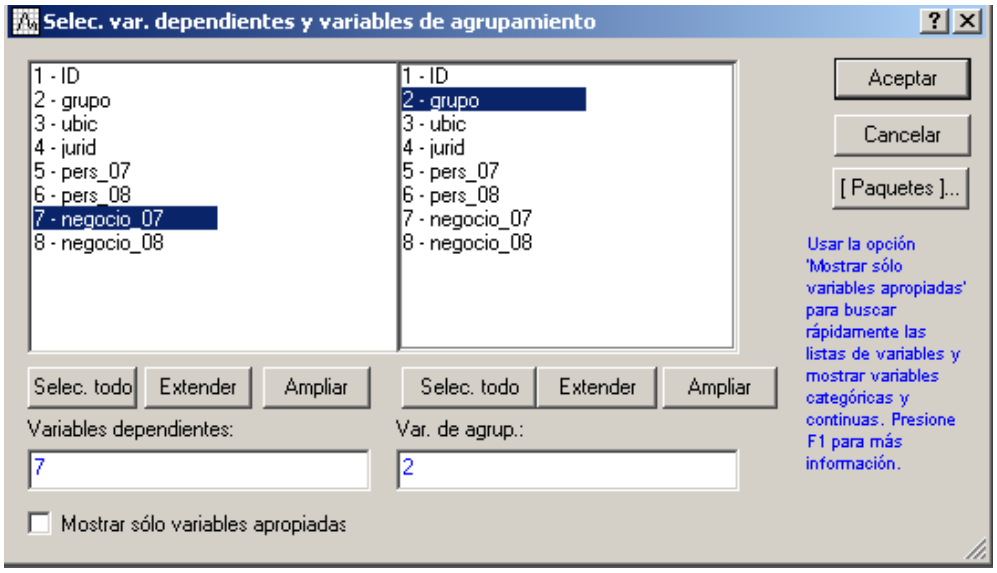
Dado que nuestro objetivo no es entretenernos con los desarrollos teóricos, vamos a ver, sobre el ejemplo mencionado, cómo solicitar al *Statistica* un ANOVA, y cómo interpretarlo. Para empezar, deberán existir en el archivo de datos dos variables: la variable cuantitativa que queremos comparar **(negocio_07),** y la variable que define los grupos a comparar **(grupo).** Por razones históricas, a esta última variable se le llama **factor**, y a sus valores, que definen los grupos a comparar, se les llama **niveles** o **tratamientos**. En nuestro caso tenemos que realizar un ANOVA de un factor con cuatro niveles o tratamientos. El procedimiento a seguir será el siguiente:

1) Ejecutamos el comando **Estadísticas>Estadísticas/Tablas básicas**. Se abre la ventana **Estadísticas básicas y tablas**



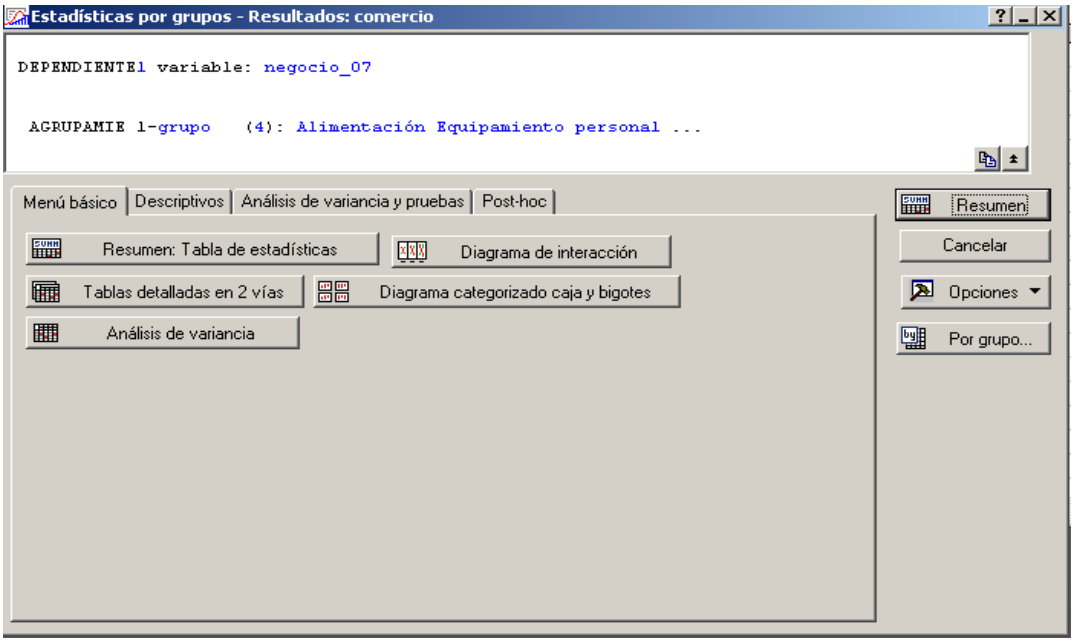
2) Seleccionamos el comando **Análisis de variancia 'Breakdown' & de una vía**. Se abre la ventana **Estadística por grupos**

3) Seleccionamos las variables que intervienen en el análisis. Para ello, en la pestaña **Tablas individuales**, pulsamos el botón **Variables**. Se abre la ventana **Selec. var. dependientes y variables de agrupamiento**.



En la lista de la izquierda (Variables dependientes) seleccionamos la variable que contiene los datos (la 7: **negocio_07**). En la lista de la derecha (Var. de agrup.) seleccionamos la variable que define los grupos (la 2: **grupo**). Pulsamos **Aceptar**. Regresamos a la ventana anterior en la que ahora aparecen la variables que acabamos de seleccionar. Mediante el botón **Códigos para las variables de agrupamiento** podemos seleccionar los grupos que nos interese comparar. Por defecto el programa los tomará todos.

4) Pulsamos **Aceptar** para ejecutar el análisis. Se abre la ventana **Estadísticas por grupos - Resultados**.



5) En la pestaña **Menú básico**, pulsamos el botón **Análisis de variancia**. Aparece en nuestro libro de trabajo una nueva hoja con los resultados esenciales del análisis:

	Análisis de variancia (comercio) Efectos marcados son significantes con p < ,05000							
Variable	**Sum.de Efecto**	gl Efecto	Med.cu Efecto	Sum.de Error	gl Error	Med.cu Error	F	p
negocio_07	45662,09	3	15220,70	1023752	72	14218,78	1,070464	0,367105

No es necesario que comprendamos todos los elementos que aparecen en esta tabla (nos ahorramos los detalles técnicos). Lo importante es que comprendamos que estamos realizando un contraste de hipótesis en el que:

> La hipótesis nula, H_0, establece que "*el volumen de negocio NO depende del grupo de actividad*", y

> La hipótesis alternativa, H_1, establece que "*el volumen de negocio SÍ depende del grupo de actividad*".

En un ANOVA interpretamos los datos de la variable cuantitativa como el resultado de extraer una muestra aleatoria de cada una de las poblaciones definidas por los niveles del factor. Es decir, ahora no estamos contemplando los valores de la variable **negocio_07** como una única muestra aleatoria de 76 comercios, sino como cuatro muestras de cuatro poblaciones: una de 37 comercios de la población "alimentación", otra de 15 comercios de la población "equipamiento personal", otra de 10 comercios de la población "equipamiento del hogar", y otra de 14 comercios de la población "otros."

Afirmar que el volumen de negocio no depende del grupo de actividad del comercio significa que pertenecer a un grupo u otro no supone ninguna diferencia para el volumen de negocio, es decir, que el volumen de negocio **es el mismo** (**en media**) en los cuatro grupos. Por tanto, la **hipótesis nula** establece la igualdad de las cuatro medias poblacionales:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

(todas las medias poblacionales son iguales)

Por su parte, afirmar que el volumen de negocio sí depende del grupo de actividad significa que pertenecer a un grupo u otro puede suponer una diferencia en el volumen de negocio, es decir, que el volumen de negocio **no es el mismo** (**en media**) en los cuatro grupos. Por tanto, la **hipótesis alternativa** establece la no igualdad de las cuatro medias poblacionales:

H_1: Existen i, j ($i \neq j$), tales que $\mu_i \neq \mu_j$ (i, j = 1,...,4)

(no todas las medias poblacionales son iguales)

Esto puede parecer un poco confuso, pero basta recordar que lo contrario de "todas" es "no todas", que no es lo mismo que "ninguna". Es decir, "no todas iguales" no significa "todas distintas." Simplemente H_1 es la negación de H_0.

Como ya sabemos de lecciones anteriores, la hipótesis nula debe ser aceptada si el *p*-valor del contraste es mayor que el nivel de significación (α), y rechazada en caso contrario. El *p*-valor aparece dentro de nuestra salida de resultados en la última columna (*p*), y vale con tres decimales p = 0'367. El nivel de significación lo elige el investigador. La opción habitual es tomar α = 0'05. Por tanto, como $p > \alpha$,

Debemos aceptar la hipótesis nula y concluir por tanto que el volumen de negocio en el año 2007 no depende del grupo de actividad

Una manera equivalente de expresar esta conclusión sería afirmar que **no existen diferencias significativas entre las cuatro medias de grupo para la variable volumen de negocio en el año 2007**. Estas medias de grupo pueden ser visualizadas fácilmente, como parte de la salida de resultados, activando la opción correspondiente, a saber:

1) Reabrimos la ventana del análisis que tenemos minimizada abajo a la izquierda.

2) Pulsamos el botón **Resumen: Tabla de estadísticas**.

3) Aparece en nuestro libro de trabajo una nueva hoja con las medias, los tamaños de grupo y las desviaciones típicas:

Tabla 'Breakdown' de estadísticas descriptivas (comercio)
N=76 (Sin datos faltantes en las lista

grupo	negocio_07 Medias	negocio_07 N	negocio_07 Desv. es
Alimentación	297,9514	37	106,5595
Equipamiento personal	322,6393	15	142,2374
Equipamiento del hogar	334,5360	10	119,6613
Otros	362,6550	14	124,9197
Todos los	319,5568	76	119,4105

A nivel descriptivo, observamos, por ejemplo, que el grupo "Otros" es el que posee un volumen de negocio medio en 2007 más elevado, y que el grupo "Alimentación" es el único con una media por debajo de la global. Sin embargo, hemos comprobado que tales diferencias no son estadísticamente significativas.

Condiciones de aplicación

Para que un Análisis de la Varianza se pueda realizar de forma válida deben cumplirse las siguientes condiciones teóricas:

1) Las muestras que comparamos deben ser aleatorias (supuesto de aleatoriedad).

2) Las muestras que comparamos deben proceder de poblaciones independientes (supuesto de independencia)

3) Las muestras que comparamos deben proceder de poblaciones normales (supuesto de normalidad).

4) Las varianzas poblacionales deben ser iguales (supuesto de homoscedasticidad).

En resumen, las muestras que comparamos tienen que ser aleatorias y deben provenir de poblaciones normales independientes que a lo sumo difieren en sus medias (precisamente la igualdad de medias es la hipótesis que sometemos a contraste).

EJERCICIO 4.3.- Estudia las condiciones de normalidad y homoscedasticidad en el análisis de varianza que acabamos de realizar.

Pruebas a posteriori (post hoc)

Si el resultado de un ANOVA fuese significativo concluiríamos que los grupos que estamos comparando no son todos iguales, que alguna diferencia existe entre ellos. Entonces se nos plantea la cuestión de averiguar qué grupos concretos son los que dan lugar a las diferencias globales detectadas. Esto implica realizar comparaciones entre parejas de grupos y por tanto nos devuelve otra vez al problema de las comparaciones múltiples. Sin embargo, se han desarrollado métodos para comparar grupos dos a dos **a posteriori**, es decir, después de que un ANOVA ha resultado significativo, de manera que el nivel de significación global se mantiene controlado en niveles aceptables.

Explicar estas pruebas a posteriori se sale del alcance (en tiempo y nivel) que hemos pensado para este curso, así que diremos simplemente que se accede a ellas en el *Statistica* mediante la pestaña **Post-hoc** de la ventana **Estadísticas por grupos - Resultados**. Si uno no está interesado en los detalles técnicos basta con que busque en la salida de resultados los *p*-valores correspondientes a la comparación de cada pareja de grupos. De todas maneras, cada prueba tiene sus ventajas e inconvenientes y pueden

no ser aplicables en cualquier circunstancia. Si no nos sentimos muy cómodos con ellas, lo mejor es consultar a un estadístico.

EJERCICIO 4.4.- Analiza si el volumen de negocio en el año 2008 depende de la naturaleza jurídica de la empresa. Estudia las condiciones de normalidad y homoscedasticidad.

EJERCICIO 4.5.- Analiza si el volumen de negocio en el año 2007 depende de la ubicación. Comprueba que los resultados coinciden con los obtenidos mediante la prueba t para dos muestras. Estudia las condiciones de normalidad y homoscedasticidad.

EJERCICIO 4.6.- Analiza si el volumen de negocio en el año 2008 depende de la combinación grupo/ubicación (8 posibilidades). Estudia las condiciones de normalidad y homoscedasticidad.

(Indicación: Empieza por añadir una nueva variable al fichero que informe de la combinación grupo/ubicación que corresponde a cada comercio).

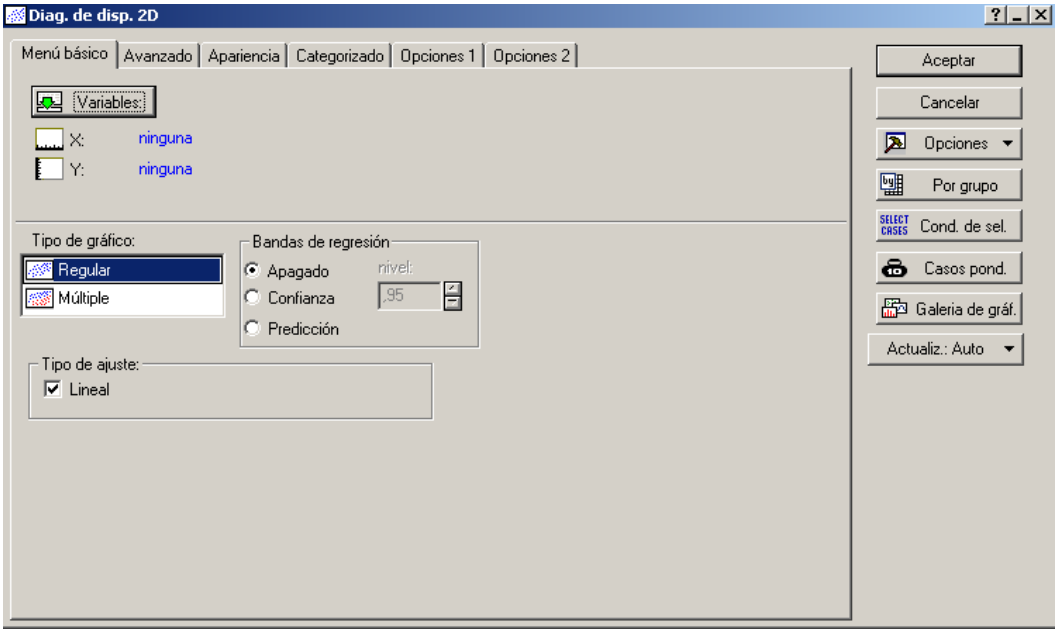
4.2. Regresión Lineal Simple

En este apartado vamos a estudiar la técnica llamada de Regresión Lineal Simple, que sirve para modelizar la relación entre dos variables cuantitativas. Consideraremos nuestro análisis en sus aspectos descriptivo e inferencial.

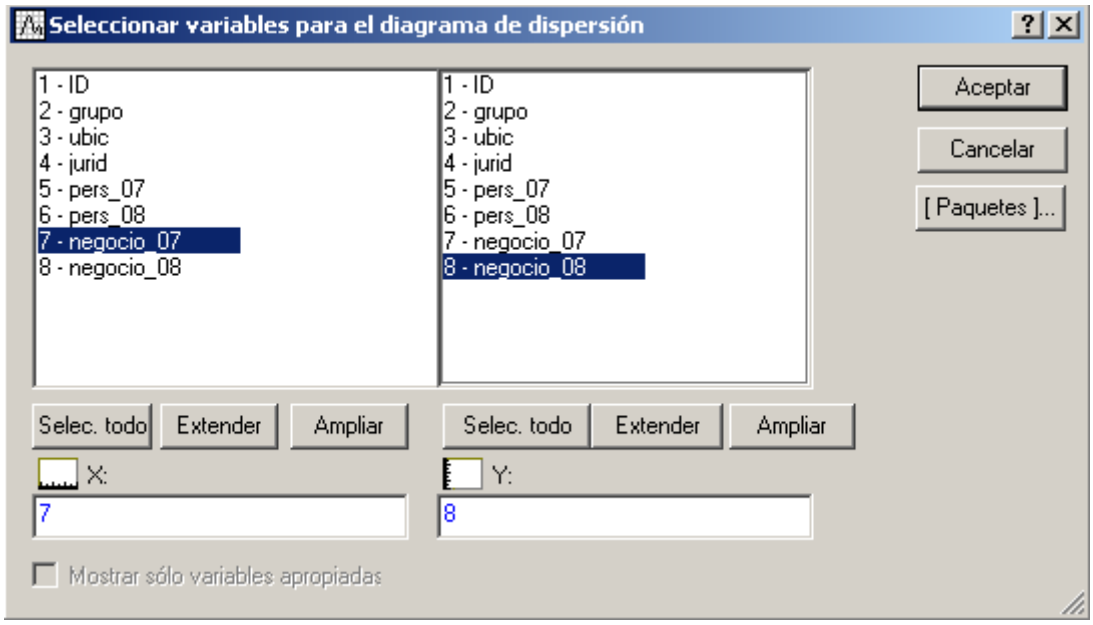
Regresión Lineal Simple: Aspecto Descriptivo

Supongamos que estamos interesados en investigar la relación que eventualmente pueda existir entre el volumen de negocio en los años 2007 y 2008, que son dos variables cuantitativas (**negocio_07** y **negocio_08**). El primer paso de nuestra aproximación al problema será construir un gráfico que nos permita hacernos una idea del tipo de relación que existe entre estas variables. El gráfico adecuado para representar datos bidimensionales cuantitativos es el llamado "diagrama de dispersión." Para construir un diagrama de dispersión con *Statistica* seguiremos los pasos siguientes:

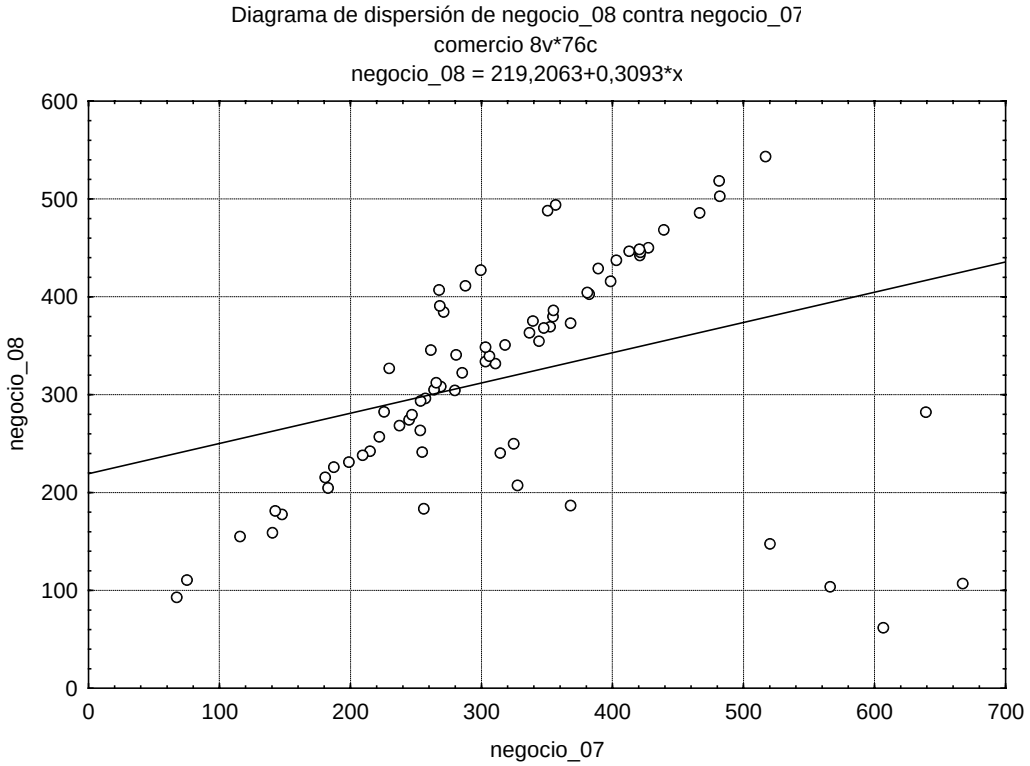
1) Ejecutamos el comando **Gráficos>Diagramas de dispersión**. Se abre la ventana **Diag. de disp. 2D**:



2) Seleccionamos las variables que contienen los datos a representar. Para ello pulsamos el botón **Variables**. Se abre la ventana **Seleccionar variables para el diagrama de dispersión**:



3) De la lista de la izquierda seleccionamos la variable que queremos representar en el eje de abscisas (la 7: **negocio_07**). De la lista de la izquierda seleccionamos la variable que queremos representar en el eje de ordenadas (la 8: **negocio_08**). Al pulsar en **Aceptar** regresamos a la ventana anterior, y pulsando **Aceptar** otra vez, *Statistica* añadirá una nueva hoja a nuestro libro de trabajo con el gráfico solicitado:



Se observa que existe una clara relación lineal, sólo distorsionada por unos pocos comercios que se desvían llamativamente del patrón general. El programa ha incluido automáticamente en el gráfico una recta de regresión (su ecuación forma parte del título), de la que hablaremos más adelante.

El segundo paso del análisis consiste en estudiar, ahora analíticamente, si existe relación entre estas variables. Para ello vamos a calcular la **covarianza** entre **negocio_07** y **negocio_08**.

La covarianza entre dos variables, *x* e *y*, viene dada por la expresión:

$$s_{xy} = Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

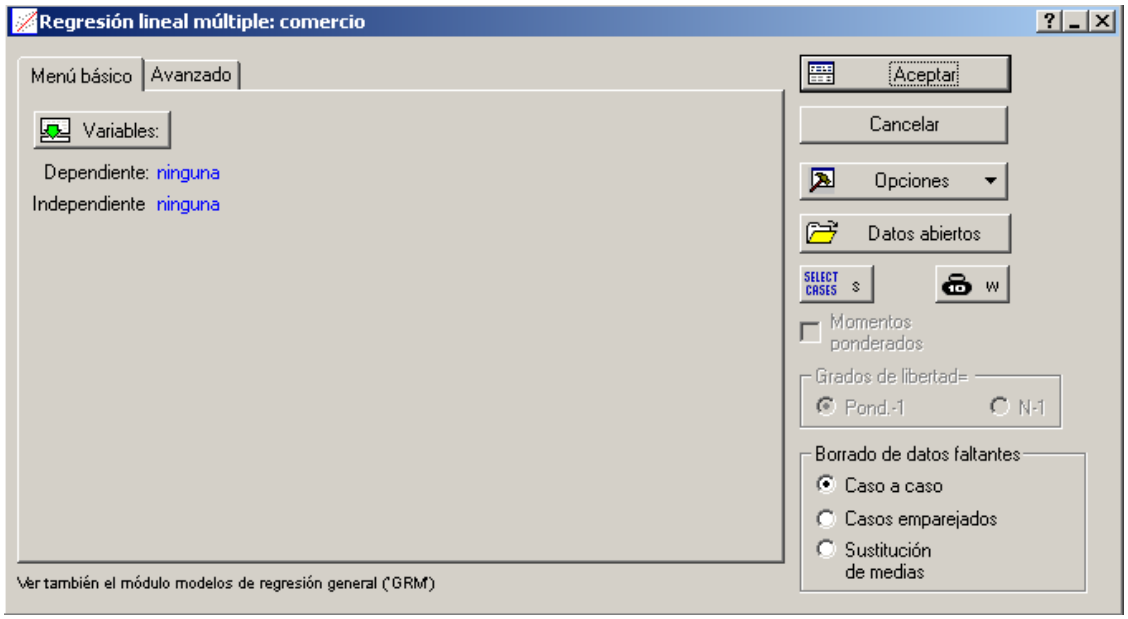
o bien, equivalentemente,

$$s_{xy} = Cov(x,y) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}\right]$$

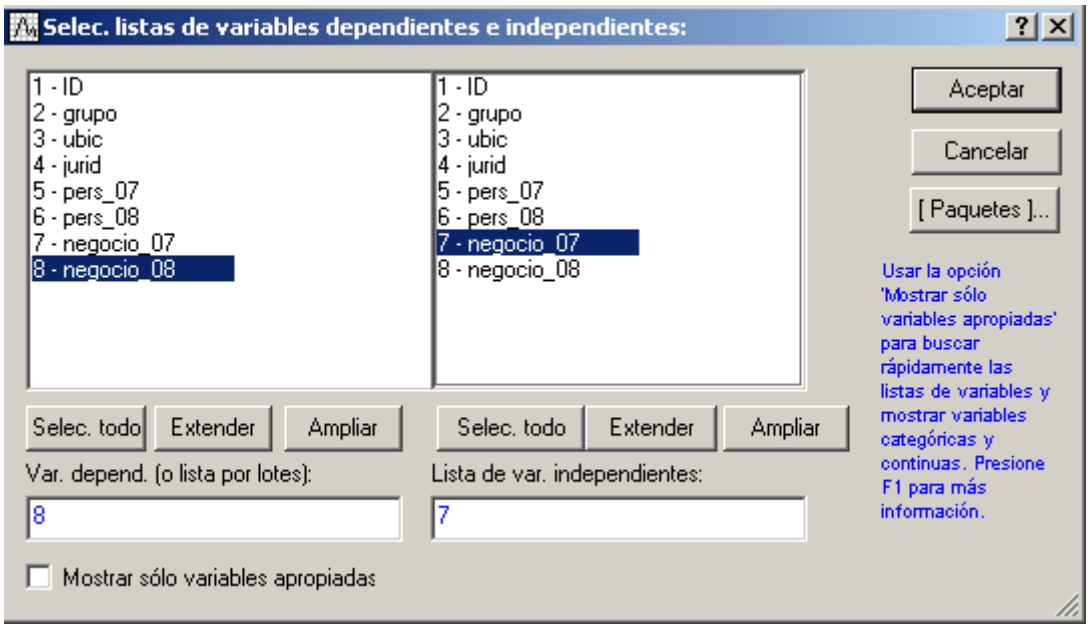
expresión que resulta más cómoda para el cálculo "a mano."

Para obtener la covarianza con *Statistica* seguiremos los pasos siguientes:

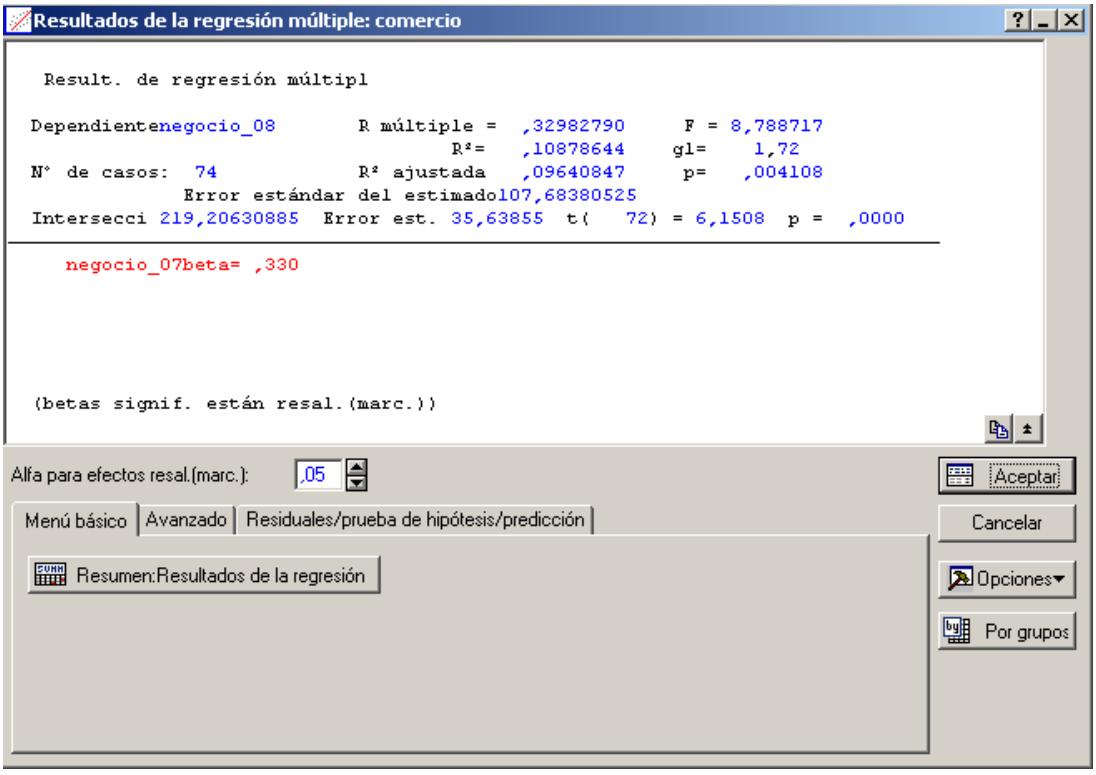
1) Ejecutamos el comando **Estadísticas>Regresión múltiple**. Se abre la ventana **Regresión lineal múltiple**:



2) Seleccionamos las variables cuya covarianza deseamos calcular. Para ello pulsamos el botón **Variables**. Se abre la ventana **Selec. listas de variables dependientes e independientes**:

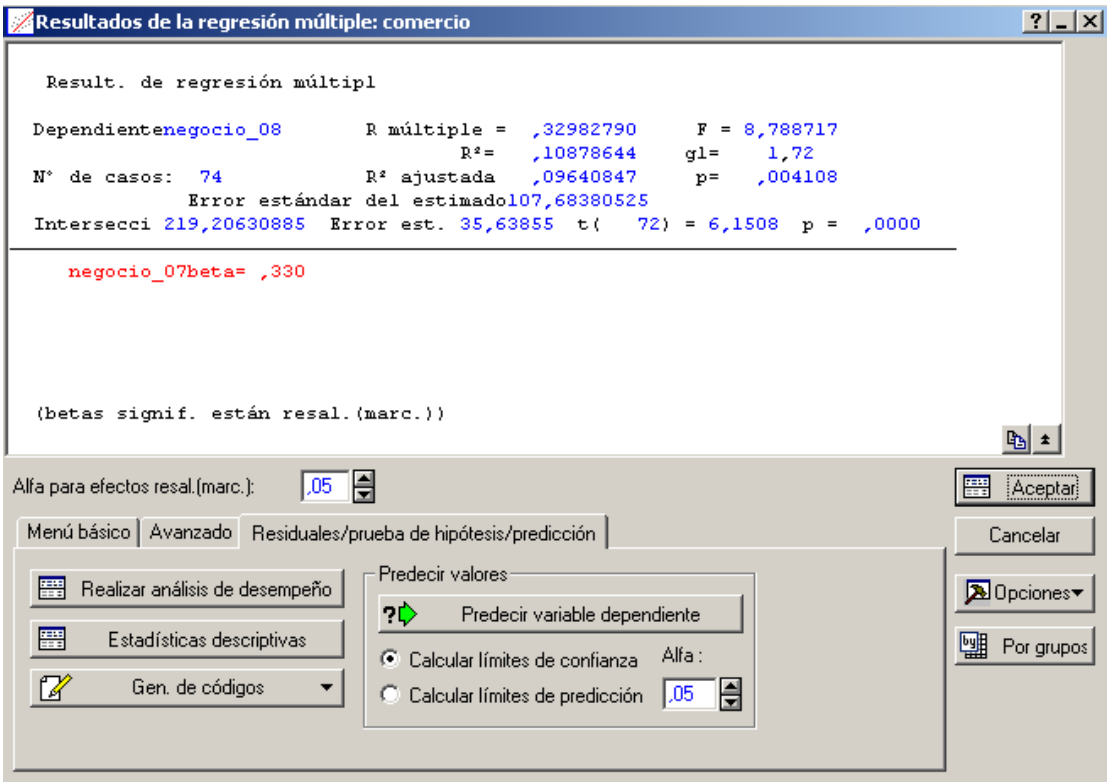


3) De la lista de la izquierda seleccionamos la variable a la que asignamos el papel de dependiente (la 8: **negocio_08**). De la lista de la derecha seleccionamos la variable a la que asignamos el papel de independiente (la 7: **negocio_07**). Al pulsar en **Aceptar** regresamos a la ventana anterior, y pulsando **Aceptar** otra vez, se abre la ventana **Resultados de la regresión múltiple**:

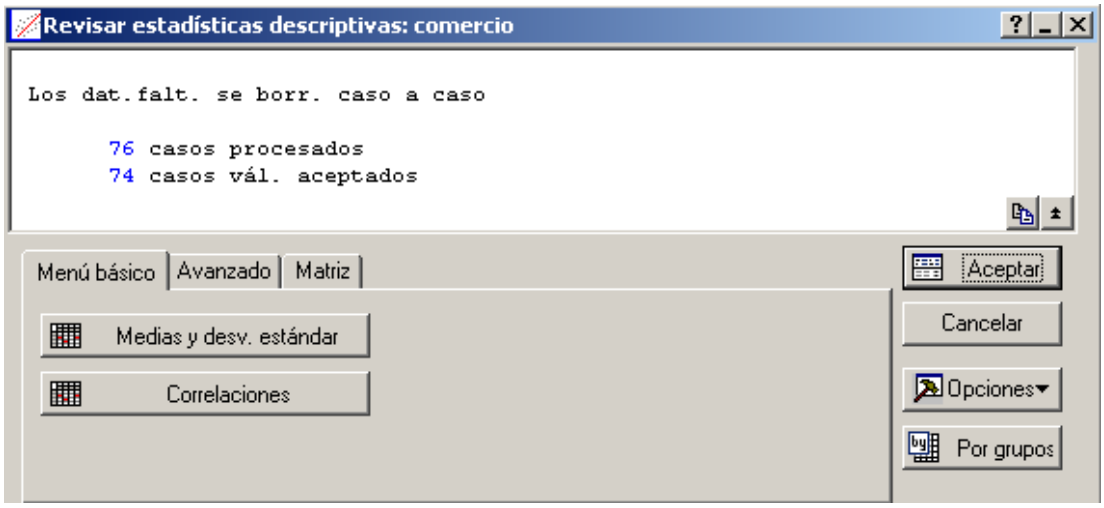

Resultados de la regresión múltiple: comercio
Result. de regresión múltipl
Dependientenegocio_08 R múltiple = ,32982790 F = 8,788717
R²= ,10878644 gl= 1,72
N° de casos: 74 R² ajustada ,09640847 p= ,004108
Error estándar del estimado107,68380525
Intersecci 219,20630885 Error est. 35,63855 t(72) = 6,1508 p = ,0000
negocio_07beta= ,330
(betas signif. están resal.(marc.))
Alfa para efectos resal.(marc.): ,05
Aceptar
Cancelar
Opciones
Por grupos
Menú básico
Avanzado
Residuales/prueba de hipótesis/predicción
Resumen:Resultados de la regresión

Esta ventana está dividida en una parte superior con un "adelanto" de los principales resultados del análisis, y una parte inferior con pestañas y botones que nos permitirán obtener los resultados que nos interesen en hojas de nuestro libro de trabajo. Veamos dónde podemos solicitar la covarianza entre nuestras variables.

4) En la pestaña **Residuales/prueba de hipótesis/predicción**, pulsamos el botón **Estadísticas descriptivas**:

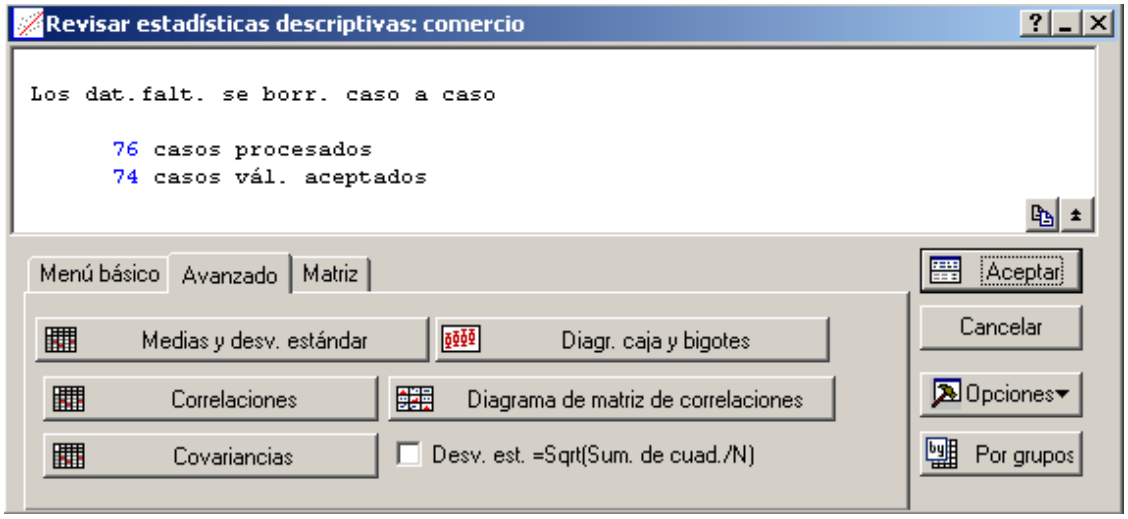


Se abre la ventana **Revisar estadísticas descriptivas**:



Statistica nos informa de que se están utilizando 74 de los 76 casos del archivo de datos (hay dos comercios para los que no consta el valor de **negocio_08**).

5) En la pestaña **Avanzado**, pulsamos el botón **Covariancias**:



Statistica añade una nueva hoja a nuestro libro de trabajo con la llamada **matriz de covarianzas**:

	Covariancias (comercio)	
Variable	**negocio_07**	negocio_08
negocio_07	14595,74	4514,03
negocio_08	4514,03	12833,01

Cada elemento de esta matriz es la covarianza entre las variables de su fila y su columna. Es sencillo comprobar a partir de la fórmula que hemos visto más arriba que la covarianza entre una variable y ella misma es su varianza, por lo que la diagonal principal de una matriz de covarianzas contiene las varianzas de cada variable. Además una matriz de covarianzas siempre será simétrica, porque *Cov(x,y) = Cov(y,x)*. Resulta, en definitiva, que la covarianza entre **negocio_07** y **negocio_08** vale 4514.03. Muy bien, y esto ¿qué significa? Para interpretar la covarianza tendremos en cuenta los siguientes puntos.

Existencia de correlación

1) La covarianza mide dependencia lineal entre las variables.
2) Si la covarianza vale 0, no existe dependencia lineal entre las variables. Se dice que las variables son **incorreladas**.
3) Si la covarianza no vale 0, existe dependencia lineal entre las variables. Se dice que las variables están **correlacionadas**.

Para nuestros datos la covarianza no vale 0, y concluimos por tanto que las variables **negocio_07** y **negocio_08** están correlacionadas: existe dependencia lineal entre ellas (cosa que ya habíamos intuido a la vista del diagrama de dispersión).

Sentido de la correlación

4) Si la covarianza es positiva, se dice que la correlación existente es **positiva**, o **directa**. Esto significa que ambas variables covarían en el mismo sentido: cuando aumenta una tiende a aumentar la otra, cuando disminuye una tiende a disminuir la otra.

5) Si la covarianza es negativa, se dice que la correlación existente es **negativa**, o **inversa**. Esto significa que ambas variables covarían en sentidos opuestos: cuando aumenta una tiende a disminuir la otra, cuando disminuye una tiende a aumentar la otra.

Para nuestros datos, la covarianza ha salido positiva, lo que indica que la correlación existente entre las variables **negocio_07** y **negocio_08** es positiva, o directa, es decir, a mayor volumen de negocio en 2007, mayor volumen de negocio en 2008 (como tendencia estadística, pueden existir casos particulares en los que haya sucedido lo contrario).

Fuerza de la correlación

Una vez que hemos descubierto que existe correlación lineal entre dos variables, y su sentido positivo o negativo, interesa valorar si tal correlación es fuerte o débil, es decir si la dependencia lineal es estrecha o más bien difusa. Esta cuestión se puede deducir intuitivamente a partir del diagrama de dispersión, que tenderá a mostrar un patrón lineal tanto más nítido cuanto más fuerte sea la correlación existente. Analíticamente, la covarianza tiende a tomar un valor "grande" (en valor absoluto) cuando la correlación es fuerte, y "pequeño" cuando la correlación es débil. Sin embargo, valorar la magnitud de una covarianza no es una tarea sencilla, debido a:

a) La covarianza "no está acotada", puede tomar cualquier valor real. Entonces ¿cómo concretar lo que significa grande o pequeño?

b) La covarianza depende de las escalas de medida de las variables. Esto significa que podemos hacer que la covarianza sea tan grande o pequeña como queramos sin más que multiplicar los datos por una constante apropiada. Esto es un contratiempo: la fuerza de la correlación entre dos variables no debería depender de la unidad de medida que se utilice para expresar los datos.

Estos problemas se resuelven mediante el **coeficiente de correlación lineal**, que viene dado por la expresión:

$$r = \frac{s_{xy}}{s_x s_y}$$

donde s_{xy} es la covarianza entre x e y
s_x es la desviación típica de x
s_y es la desviación típica de y

Las dos propiedades básicas del coeficiente de correlación son:

1) r está comprendido entre –1 y 1, ambos inclusive: $-1 \leq r \leq 1$

2) r es adimensional, en particular es independiente de las unidades de medida en las que se expresen los datos. Dicho de otra forma, es invariante frente a cambios de origen y escala en los datos.

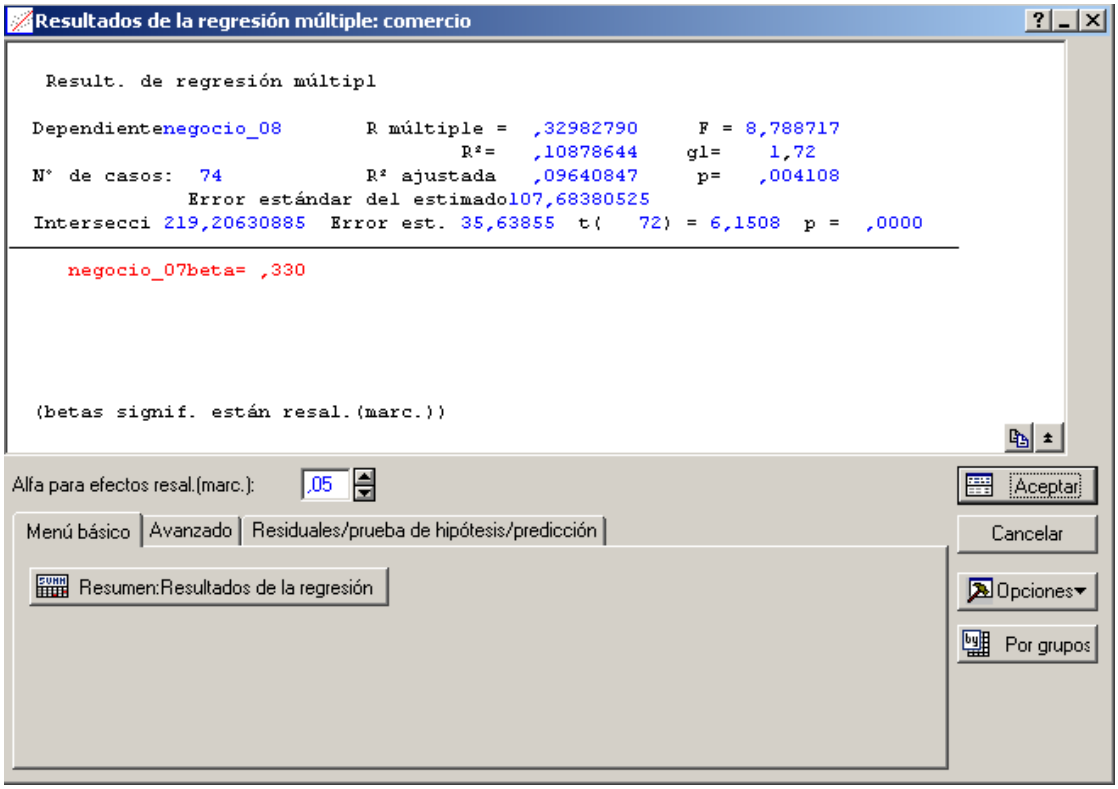
Interpretación del valor de r

1) $r = -1$ Correlación negativa exacta (los datos están sobre una recta de pendiente negativa).
2) $r \approx -1$ Fuerte correlación negativa.
3) $r \approx 0$, $r < 0$ Débil correlación negativa.
4) $r = 0$ Incorrelación.
5) $r \approx 0$, $r > 0$ Débil correlación positiva.
6) $r \approx 1$ Fuerte correlación positiva.
7) $r = 1$ Correlación positiva exacta (los datos están sobre una recta de pendiente positiva).

Los casos 1, 4 y 7 corresponden a situaciones límite que no suelen darse nunca en la práctica con datos reales. Respecto a los casos restantes, suele considerarse débil una correlación por debajo de 0.7 (**$|r| < 0.7$**), pero cualquier regla de este tipo siempre es arbitraria.

Para obtener el coeficiente de correlación con *Statistica* seguiremos los siguientes pasos:

1) Reabrimos la ventana del análisis que tenemos minimizada abajo a la izquierda. Aparecerá abierta la última ventana que hayamos visitado: **Revisar estadísticas descriptivas**. Pulsamos el botón **Cancelar** para regresar a la ventana anterior: **Resultados de la regresión múltiple**:


Resultados de la regresión múltiple: comercio
Result. de regresión múltipl
Dependientenegocio_08 R múltiple = ,32982790 F = 8,788717
R²= ,10878644 gl= 1,72
N° de casos: 74 R² ajustada ,09640847 p= ,004108
Error estándar del estimado107,68380525
Intersecci 219,20630885 Error est. 35,63855 t(72) = 6,1508 p = ,0000
negocio_07beta= ,330
(betas signif. están resal.(marc.))
Alfa para efectos resal.(marc.): ,05
Aceptar
Menú básico
Avanzado
Residuales/prueba de hipótesis/predicción
Cancelar
Resumen:Resultados de la regresión
Opciones
Por grupos

En esta ventana ya nos aparece el coeficiente de correlación (**R múltiple = 0.32982790**), pero es mejor que lo incorporemos a nuestro libro de trabajo para poder guardarlo y recuperarlo sin tener que repetir el análisis para visualizar esta ventana. Para ello:

2) En la pestaña **Menú básico**, pulsamos el botón **Resumen: Resultados de la regresión**. *Statistica* añade dos nuevas hojas a nuestro libro de trabajo. La primera de ellas es:

Estadística	Resumen estadísticas; Vanegocio_08ependiente: (comercio) Valor
R múltiple	0,329827895
R² múltiple	0,108786441
R² ajustado	0,0964084745
F(1,72)	8,78871696
p	0,00410789764
Err.est. de estimado	107,683805

Para nuestros datos, el coeficiente de correlación obtenido (en la línea **R múltiple**) es, con dos decimales, ***r* = 0.33**, lo que significa que entre las variables **negocio_07** y **negocio_08 existe una débil correlación positiva**.

El programa proporciona el valor del coeficiente de correlación en valor absoluto, por lo que hay que determinar el signo de la correlación por otros medios. Esto no supone ningún problema. En el primer paso del análisis, en el que hemos obtenido el diagrama de dispersión, la pendiente de la recta de ajuste que aparece superpuesta a los puntos de datos nos indica el sentido de la correlación.

Por supuesto, es bastante claro que para saber si existe correlación entre dos variables, si es positiva o negativa, y si es fuerte o débil, basta con calcular el coeficiente de correlación, de manera que el proceso que hemos seguido, partiendo de la covarianza, se debe más a razones didácticas que a razones prácticas.

Coeficiente de Determinación Lineal

Alternativamente, es posible utilizar, para medir la fuerza de la correlación, el llamado **coeficiente de determinación lineal**. Este coeficiente no es más que el cuadrado del coeficiente de correlación, y, por tanto, vendrá dado por la expresión:

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$$

Es evidente que se cumplirá: $0 \le r^2 \le 1$.

Interpretación del valor de r^2

1) $r^2 = 0$ Incorrelación.
2) $r^2 \approx 0$ Débil correlación.
6) $r^2 \approx 1$ Fuerte correlación.
7) $r^2 = 1$ Correlación exacta (los datos están sobre una recta).

Statistica proporciona el valor de r^2 junto con el de r (en la línea **R^2 múltiple**). Observamos, pues, que para nuestros datos se tiene $\boldsymbol{r^2 = 0.11}$. En términos de r^2, el criterio para considerar débil una correlación ($|r| < 0.7$), se convierte en $\boldsymbol{r^2 < 0.5}$ (tomando el valor redondo 0.5 para el cuadrado de 0.7). Volveríamos a concluir que entre las variables **negocio_07** y **negocio_08** **existe un débil correlación** (positiva, desde luego, como sabemos desde que hemos obtenido el diagrama de dispersión).

El coeficiente de determinación tiene además una interesante interpretación: expresa en qué medida la variable *x* determina (linealmente) el valor de *y*. El valor $r^2 = 0.11$ (11%, si lo expresamos en porcentaje) nos está diciendo que el volumen de negocio en 2007 determina en un 11% el volumen de negocio en 2008. ¿Cómo entender esta afirmación? Aceptamos que el valor en el 2008 depende del valor en el 2007, ¡pero sólo hasta cierto punto! Existen multitud de factores (conocidos o desconocidos, controlables o incontrolables) que determinan el volumen de negocio de un año particular, entre los cuales el volumen en el año anterior no es más que uno de tantos, ¿o es especialmente importante? No lo parece, porque el peso relativo de su influencia no alcanza más que el 11%, que es bastante poco, dejando un amplio margen del 89% para la influencia de todos los demás factores. Técnicamente, el coeficiente de determinación mide la proporción de la varianza de *y* "explicada" por la relación que existe con *x*. El criterio $r^2 < 0.5$ para considerar débil una correlación se basa en exigir que una variable determine el valor de la otra por lo menos en un 50%, de manera que quede un margen inferior a dicho 50% para la influencia del resto de factores posibles (que para nosotros representa la componente aleatoria de la relación entre las variables).

EJERCICIO 4.7.- Para las variables **negocio_07** y **negocio_08**, construye el diagrama de dispersión, calcula e interpreta la covarianza, el coeficiente de correlación y el coeficiente de determinación, eliminando del análisis los cinco comercios que se desvían del patrón general marcado por el resto. Compara los resultados con los que hemos obtenido usando todos los datos.

Modelización de la relación entre las variables: Rectas de Regresión.

El siguiente paso a dar en el análisis es el de la construcción de un modelo matemático que represente la relación entre nuestras variables. Supongamos que estamos interesados en modelizar la variable **negocio_08** como función lineal de la variable **negocio_07**. Esto significa que queremos construir una expresión de la forma:

$$y = a + b\,x$$

siendo x = negocio_07
y = negocio_08
a, b = parámetros a determinar a partir de los datos disponibles para x e y.

Esta expresión matemática corresponde a la ecuación de una recta con pendiente *b* y ordenada en el origen *a*. Las fórmulas para calcular *a* y *b* son las siguientes:

$$b = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

(Obsérvese que estas fórmulas están pensadas para calcular primero *b* y luego llevar el valor obtenido a la fórmula de *a*).

Statistica proporciona los valores de *a* y *b* en la segunda de las dos ventanas que acabamos de incorporar al libro de trabajo:

Resumen de la regresión para variable dependiente: negocio_08 (comercio)
R = ,32982790 R²= ,10878644 R² ajustado = ,09640847
F(1,72)=8,7887 p<,00411 Error est. de la estima107,68

N=74	Beta	Err.est. de Beta	B	Err.est. de B	t(72)	nivel-p
Intersección			219,2063	35,63856	6,150819	0,000000
negocio_07	0,329828	0,111256	0,3093	0,10432	2,964577	0,004108

En la línea **Intersección**, columna **B**, aparece la ordenada en el origen **a = 219.2063**, y en la línea **negocio_07**, columna **B**, aparece la pendiente **b = 0.3093**. La ecuación de la recta que modeliza la variable **negocio_08** como función lineal de la variable **negocio_07** es, pues:

negocio_08 = 219.2063 + 0.3093 negocio_07

Esta recta (que *Statistica* ha incluido automáticamente en el diagrama de dispersión) recibe el nombre de "**recta de regresión mínimo-cuadrática de *y* sobre *x***." La denominación mínimo-cuadrática se debe a que el método que se ha utilizado para construirla, y del que derivan las fórmulas de *b* y *a* anteriores, es el "método de los mínimos cuadrados." La denominación "*y* sobre *x*" hace referencia a los papeles asignados a las variables que se manejan. El modelo se ha construido asignando a *x* (**negocio_07**) el papel de variable independiente (o explicativa), y a *y* (**negocio_08**) el papel de variable dependiente (o respuesta). Si intercambiamos estos papeles obtendremos una recta distinta, que será de la forma:

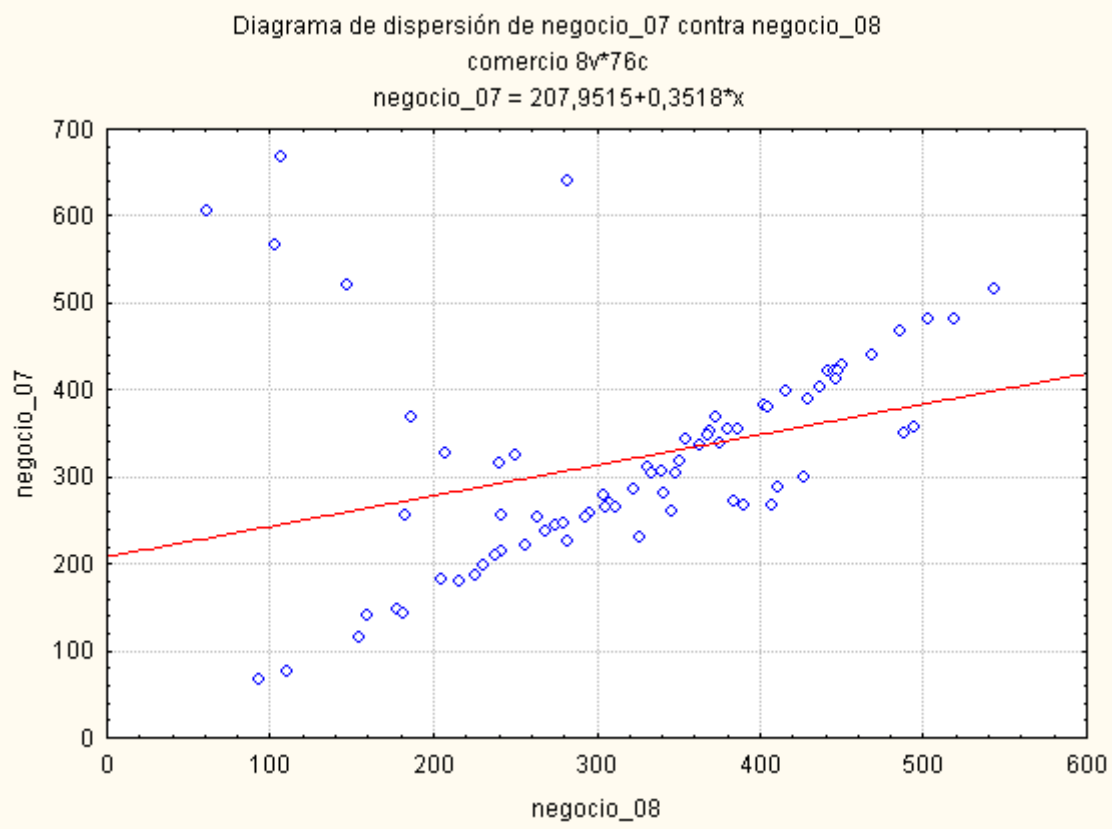
$$x = a' + b'\,y$$

Las fórmulas para *b'* y *a'* serán:

$$b' = \frac{s_{xy}}{s_y^2}$$

$$a' = \bar{x} - b'\bar{y}$$

Si estamos interesados en modelizar **negocio_07** en términos de **negocio_08**, podemos empezar por construir el diagrama de dispersión:



y, a continuación, obtener los valores de los parámetros del modelo (aunque ya aparecen en el título del gráfico):

Resumen de la regresión para variable dependiente: negocio_07 (comercio)
R = ,32982790 R²= ,10878644 R² ajustado = ,09640847
F(1,72)=8,7887 p<,00411 Error est. de la estima114,84

N=74	Beta	Err.est. de Beta	B	Err.est. de B	t(72)	nivel-p
Intersección			207,9515	40,03759	5,193906	0,000002
negocio_08	0,329828	0,111256	0,3518	0,11865	2,964577	0,004108

de manera que la ecuación de la recta que modeliza la variable **negocio_07** como función lineal de la variable **negocio_08** es:

negocio_07 = 207.9515 + 0.3518 negocio_08

Esta recta recibe el nombre de "**recta de regresión mínimo-cuadrática de x sobre y**."

Resulta, pues, que para unos datos bidimensionales cuantitativos, no existe una única recta de regresión, sino dos. Cuál utilizar depende de la aplicación que queramos hacer del modelo. La recta de *y* sobre *x* es la óptima (es decir, la mejor recta posible, en el sentido de los mínimos cuadrados) para predecir el valor de *y* a partir de un valor dado de *x*, y la recta de *x* sobre *y* es la óptima para predecir el valor de *x* a partir de un valor dado de *y*. Por otra parte, en la recta de *y* sobre *x*, el parámetro *b* representa el efecto lineal que tiene la variable *x* en la variable *y*, mientras que, en la recta de *x* sobre *y*, el parámetro *b'* representa el efecto lineal que tiene la variable *y* en la variable *x*.

EJERCICIO 4.8.- Para las variables **negocio_07** y **negocio_08**, calcula las rectas de regresión de "y sobre x" y de "x sobre y", eliminando del análisis los cinco comercios "atípicos". Compara los gráficos en los que aparecen estas rectas con los obtenidos usando todos los datos.

Regresión Lineal Simple: Aspecto Inferencial

Generalmente, los datos que se utilizan para estudiar y modelizar la relación existente entre dos variables cuantitativas corresponden a una muestra de individuos pertenecientes a cierta población. Se plantea entonces la cuestión, propia de la Inferencia Estadística, de tratar de alcanzar conclusiones generales para toda la población a partir de la información proporcionada por la muestra. Así, en el contexto en el que nos encontramos, no nos interesará únicamente si la correlación observada entre dos variables es positiva o negativa, fuerte o débil, sino también si **es significativa**.

El hecho de que exista correlación a nivel muestral **no implica** que tenga que existir necesariamente a nivel poblacional. Puede ser que dos variables sean incorreladas a nivel poblacional y sin embargo no lo sean en la muestra particular analizada. La muestra habrá sido (debería haber sido) seleccionada al azar, y por tanto los estadísticos que calculamos a partir de ella están sometidos a las fluctuaciones aleatorias inherentes al proceso de muestreo.

En este apartado vamos a estudiar el punto más importante del enfoque inferencial de la modelización, que es el llamado **Contraste de la Regresión**.

La recta de regresión que hemos obtenido antes, $y = a + b\,x$, es ahora la recta de regresión muestral. Denotemos la recta de regresión poblacional por $y = \alpha + \beta x$. El contraste de la regresión es un contraste de hipótesis en el que la hipótesis nula es:

$$\mathbf{H_0: \beta = 0}$$

y la hipótesis alternativa es:

$$\mathbf{H_1: \beta \neq 0}$$

De ser cierta la hipótesis nula, el modelo en la población carecería de término en *x*; la recta de regresión poblacional sería de la forma *y = cte*, recta horizontal que no representa ninguna relación entre *x* e *y*. Equivalentemente, en la población tanto la covarianza como el coeficiente de correlación se anularían (basta pensar en las fórmulas para *r* y *b* en términos poblacionales), es decir las variables *x* e *y* serían incorreladas a nivel poblacional.

Por el contrario, si la hipótesis nula es falsa (es cierta la alternativa), el modelo en la población incluiría el término en *x*; la recta de regresión poblacional tendría pendiente no nula, representando una relación lineal entre *x* e *y*. Equivalentemente, en la población tanto la covarianza como el coeficiente de correlación serían distintos de cero, es decir, las variables *x* e *y* estarían correlacionadas a nivel poblacional.

Por supuesto, es imposible saber si la hipótesis nula es cierta o falsa, pero el contraste de la regresión nos va a permitir decidir si los datos muestrales proporcionan evidencia suficiente para descartar H_0 en favor de H_1.

Statistica proporciona los resultados del contraste de la regresión junto con los valores de los coeficientes de la recta de regresión muestral en la correspondiente hoja del libro de trabajo que ya obtuvimos antes (para la regresión de y = **negocio_08** sobre x = **negocio_07**):

Resumen de la regresión para variable dependiente: negocio_08 (comercio)
R = ,32982790 R²= ,10878644 R² ajustado = ,09640847
F(1,72)=8,7887 p<,00411 Error est. de la estima107,68

N=74	Beta	Err.est. de Beta	B	Err.est. de B	t(72)	**nivel-p**
Intersección			219,2063	35,63856	6,150819	0,000000
negocio_07	0,329828	0,111256	0,3093	0,10432	2,964577	0,004108

Como sabemos, hay que rechazar la hipótesis nula, en favor de la alternativa, si el *p-valor* del contraste es suficientemente pequeño ($p < \alpha$, siendo α el nivel de significación elegido). El *p-valor* aparece en la línea **negocio_07**, columna **nivel-p**, siendo, por tanto, *p = 0.004108*. Se trata de un *p-valor* pequeño para cualquiera de los niveles de significación usados habitualmente (0.10, 0.05, 0.01), y por tanto **debemos rechazar la hipótesis nula en favor de la alternativa**, concluyendo que

en la población β es distinto de cero

Esta conclusión es equivalente a afirmar que

en la población existe correlación lineal entre las variables

negocio_07 y negocio_08

Otra manera muy típica de expresar esta misma conclusión es que

la correlación lineal <u>observada en la muestra</u> entre las variables negocio_07 y negocio_08 es <u>estadísticamente significativa</u> (al 10%, 5% ó 1%, según el nivel de significación elegido)

Y aún otra manera equivalente de expresarnos sería:

Se ha detectado <u>en la muestra</u> un efecto lineal significativo (representado por b = 0.3093) de la variable negocio_07 sobre la variable negocio_08

Desde luego, si la hipótesis nula hubiera sido aceptada, las conclusiones serían exactamente contrarias, en particular diríamos que la correlación observada en la muestra no es estadísticamente significativa.

<u>EJERCICIO 4.9</u>.- Para las variables **negocio_07** y **negocio_08**, realiza e interpreta el contraste de la regresión para las rectas de "y sobre x" y de "x sobre y", eliminando del análisis los cinco comercios "atípicos". Compara los p-valores con los obtenidos usando todos los datos.

BIBLIOGRAFÍA

Delgado de la Torre, Rosario:
Probabilidad y estadística para ciencias e ingenierías.
Delta Publicaciones. Madrid, 2008.

Martín-Pliego López, Fco.Javier:
Introducción a la estadística económica y empresarial. Teoría y práctica.
Thomson. Madrid, 2004.

Peña, Daniel:
Fundamentos de estadística.
Alianza Editorial. Madrid, 2001.

Ruiz-Maya Pérez, Luis; Martín-Pliego, F. Javier:
Fundamentos de inferencia estadística.
Thomson. Madrid, 2002.

Trabajo final:

Opción 1: Realiza un análisis estadístico con los datos contenidos en un fichero dado.

El archivo ais del paquete DAAG del software estadístico R, contiene información de 13 variables observadas en 202 atletas.
La descripción completa del fichero es la siguiente:

ais → Australian athletes data set

Description
These data were collected in a study of how data on various characteristics of the blood varied with sport body size and sex of the athlete.

Usage
data(ais)

Format
A data frame with 202 observations on the following 13 variables.
rcc red blood cell count, in $10^{12}.l^{-1}$ (billones por litro)
wcc white blood cell count, in $10^{12}.l^{-1}$ (billones por litro)
hc hematocrit, in percent
hg hemaglobin concentration, in g per decaliter (g/dl)
ferr plasma ferritins, in ng dl^{-1}
bmi Body mass index, in $kg.m^{-2}$
ssf sum of skin folds
pcBfat percent Body fat
lbm lean body mass, in kg
ht height, cm
wt weight, kg
sex a factor with levels: f , m
sport a factor with levels: B_Ball (p), Field(p), Gym(p), Netball(p), Row(e), Swim(e), T_400m(e), T_Sprnt(p), Tennis(e), W_Polo(e)

Details
Do blood hemoglobin concentrations of athletes in endurance-related events differ from those in power-related events?

Source
These data were the basis for the analyses that are reported in Telford and Cunningham (1991).

References
Telford, R.D. and Cunningham, R.B. 1991. Sex, sport and body-size dependency of hematology in highly trained athletes. Medicine and Science in Sports and Exercise 23: 788-794.

Nota: Los valores indicados entre paréntesis a continuación del nombre de cada deporte, (p) y (e), clasifican los mismos en dos categorías: p= power-related (de potencia) y e=endurance-related (de resistencia).

1) Realiza un estudio descriptivo de: sexo (**sex**), **tipo** de deporte (power/endurance), y concentración de hemoglobina (**hg**), de los atletas de la muestra.

2) Realiza un estudio descriptivo de la concentración de hemoglobina de cada uno de los grupos de atletas en los que dividen a la muestra las variables sexo y tipo (considerándolas individualmente –2 grupos por sexo y 2 grupos por tipo- y combinadamente –4 grupos por sexo+tipo-).

3) Estamos interesados en responder a las siguientes cuestiones referidas a los deportistas de Australia:
   - ¿Existen diferencias significativas en las medias de las concentraciones de hemoglobina entre los hombres y las mujeres que practican deportes de resistencia?
   - Por separado para hombres y mujeres: La concentración de hemoglobina en sangre ¿difiere entre los atletas que practican deportes de resistencia y los que practican deportes de potencia?

4) Análisis de varianza: Selecciona una variable antropométrica y una variable fisiológica y estudia si existen diferencias entre los grupos "sexo/tipo de deporte". Estudia las condiciones de validez.

5) Regresión lineal: Modeliza la relación que existe entre las variables antropométrica y fisiológica que has seleccionado para el análisis de varianza anterior (gráfico, covarianza, correlación, determinación, rectas, contraste de la regresión).

Opción 2: A partir de un fichero de datos propios, realiza un análisis estadístico en el que estés interesado. Dicho análisis ha de seguir el esquema propuesto en la Opción 1: estudio descriptivo, contraste de hipótesis, análisis de varianza y estudio de regresión.

Indicaciones:

Todas las respuestas han de estar debidamente justificadas a partir de los resultados obtenidos (tablas, gráficos, estadísticos, métodos, etc).

Hay que comentar lo que se hace y por qué se hace. Por ejemplo, la utilización de una determinada herramienta (tipo de gráfico, tipo de contraste de hipótesis, etc) ha de ser justificada, en particular, comprobando las condiciones de aplicación (o de validez).